Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1033–1043
Marseille, 11–16 May 2020

The Discussion Tracker Corpus of Collaborative Argumentation

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, Amanda Godley
University of Pittsburgh
{cao48, lul32, ras306, dlitman, agodley}@pitt.edu

Abstract

Although Natural Language Processing (NLP) research on argument mining has advanced considerably in recent years, most studies draw on corpora of asynchronous and written texts, often produced by individuals. Few published corpora of synchronous, multi-party argumentation are available. The Discussion Tracker corpus, collected in American high school English classes, is an annotated dataset of transcripts of spoken, multi-party argumentation. The corpus consists of 29 multi-party discussions of English literature transcribed from 985 minutes of audio. The transcripts were annotated for three dimensions of collaborative argumentation: argument moves (claims, evidence, and explanations), specificity (low, medium, high) and collaboration (e.g., extensions of and disagreements about others' ideas). In addition to providing descriptive statistics on the corpus, we provide performance benchmarks and associated code for predicting each dimension separately, illustrate the use of the multiple annotations in the corpus to improve performance via multi-task learning, and finally discuss other ways the corpus might be used to further NLP research.

Keywords: Corpus, Discourse, Text mining

1. Introduction

Natural Language Processing (NLP) research has advanced considerably in recent years, developing reliable predictors for argumentation (Mirkin et al., 2018; Lippi and Torroni, 2016; Aharoni et al., 2014; Biran and Rambow, 2011a; Carlile et al., 2018; Habernal et al., 2014; Park and Cardie, 2018; Stab and Gurevych, 2014; Stab and Gurevych, 2017a), specificity (Li and Nenkova, 2015; Li et al., 2016; Louis and Nenkova, 2012; Gao et al., 2019), and collaboration (Richey et al., 2016). The majority of corpora informing these developments are made up of written texts (e.g., documents, asynchronous online discussions) extracted from the web and often written by individuals. Additionally, annotations in these corpora typically focus on one linguistic dimension at a time (e.g., specificity *or* dimensions of argumentation). As such, few published corpora both (1) focus on synchronous multi-party argumentation and (2) include multiple simultaneous annotations.

To address the lack of multi-party synchronous argumentation corpora that include multiple simultaneous annotations, we are releasing the Discussion Tracker corpus, which includes transcriptions of 29 multi-party synchronous argument-based dialogues collected in high school English classes and annotated for three simultaneous but distinct discourse dimensions (argumentation, specificity and collaboration). In addition to providing descriptive statistics on the corpus, we provide code and benchmark performance figures for predicting each of the three annotated dimensions, illustrating the challenging nature of this corpus. We then illustrate the benefits the corpus has already afforded NLP algorithms for improving argument mining with multi-task learning and discuss other potential uses of the corpus for further NLP research.

2. Related Work

The development of argument mining tasks has been heavily informed by corpora of written texts extracted mostly from the web (e.g., online newspapers, Wikipedia, blog posts, user comments) (Al Khatib et al., 2018; Biran and Rambow, 2011b; Aharoni et al., 2014; Habernal et al., 2014; Park and Cardie, 2018; Swanson et al., 2015; Cabrio and Villata, 2014; Boltužić and Šnajder, 2014), from student essays (Stab and Gurevych, 2014; Stab and Gurevych, 2017a; Carlile et al., 2018), from legal documents (parliamentary records and court reports) (Mochales and Moens, 2011; Ashley and Walker, 2013), or from speeches (Mirkin et al., 2018; Lippi and Torroni, 2016). Similarly, NLP developments in specificity have typically drawn on newspaper articles (Louis and Nenkova, 2011; Li and Nenkova, 2015; Li et al., 2016) or online content (Gao et al., 2019; Ko et al., 2019). Although work in Computer Supported Collaborative Learning (CSCL) (Weinberger and Fischer, 2006; Fischer et al., 2013; Noroozi et al., 2013; Scheuer et al., 2014; Dillenbourg and Hong, 2008) has used mainly multi-party online data, it has typically drawn from asynchronous written discourse as opposed to synchronous dialogue.

Many corpora of written text data that have been annotated for similar dimensions as the Discussion Tracker corpus have focused on single and independent tasks like argument components (e.g. claims, premises) (Biran and Rambow, 2011b; Aharoni et al., 2014; Habernal et al., 2014; Mochales and Moens, 2011), relations between pairs of arguments (e.g. support, attack) (Park and Cardie, 2018), or specificity (Li and Nenkova, 2015; Li et al., 2016; Louis and Nenkova, 2012). Some corpora include multi-level annotations that allow for performing multiple tasks (e.g. argument components and relations) (Al Khatib et al., 2018; Stab and Gurevych, 2014; Stab and Gurevych, 2017a; Ashley and Walker, 2013; Carlile et al., 2018). However, multi-level annotations like these have often been heavily dependent on one another. For example, although Carlile et al. analyzed two simultaneous annotations (argument components and specificity), their definition of specificity was contingent on argument component type, defining specificity of claims differently from specificity of premises. In this way, their multi-level annotations were tightly coupled and could only be analyzed in conjunction with one an-

other. In cases like Al Khatib et al. (2018), which provides distinct multi-level annotations for asynchronous online argumentation, they do not include specificity at all. By providing annotations for multi-level and distinct annotations for collaboration, argumentation, and specificity in synchronous dialogues in natural learning environments, we believe the Discussion Tracker corpus is capable of offering the NLP community new opportunities for research.

Of the extant multi-party synchronous data, the most similar corpus to ours is Richey et al.'s SRI corpus (Richey et al., 2016). Their corpus was developed for analysis of the ways student talk patterns correlated with collaborative learning. Equipped with multi-speaker, small group audio recordings of middle school students discussing mathematical solutions, the SRI corpus provides multi-level annotations of collaboration indicators and collaboration quality. However, although the SRI corpus includes multi-party synchronous argumentation data, some major differences stand out from the Discussion Tracker corpus. First, the scope of the the Discussion Tracker corpus extends beyond collaborative dimensions to include dimensions of argumentation and specificity as well, whereas the SRI corpus focuses only on collaboration. Second, whereas the group size of the multi-party dialogues in the SRI corpus is three students, the group sizes in the Discussion Tracker corpus average around 15 students per discussion. Third, in addition to providing the gender of speakers, as the SRI corpus does, the Discussion Tracker corpus also includes identification of the racial background of speakers. Fourth, Discussion Tracker will be released as written transcriptions with corresponding annotations whereas the SRI data is released as audio files with time-stamped annotations. The differences between these corpora are not to suggest one is better than the other, but simply that different formats afford different avenues for analysis.

3. Data Collection

The Discussion Tracker corpus is based on audio-recorded multi-party spoken discussions in 10 different high school English teachers' classrooms across three different school districts (suburban, urban, and rural) (see Table 1). Between October 2018 - March 2019 we recorded a total of three literature discussions per classroom. Omitting one discussion that was off-topic, the corpus we are releasing contains 29 transcriptions of high school literature discussions based on 985 minutes of audio (see Table 2).

Across the ten classrooms, the mean number of discussion participants was 15 students (SD 6), ranging from 6 to 29. In accordance with educational research examining racial and gender inequities in instructional practice (Kelly, 2008; Sherry, 2014), we collected metadata for race and gender demographics. Based on notes taken during data collection, we estimated that on average student discussants were 50% male and 50% female (SD 0.18), 77% white (SD 0.22) and 23% nonwhite (SD 0.22). Of the nonwhite students most appeared to be Indian (58%), 18% appeared Black, 15% East Asian, and the remaining 9% appeared Latinx or other.

In order to maintain the authenticity of the instructional environment, teachers were free to facilitate their discussions according to their pedagogical expertise, so long as they arranged students to face the microphone (which was placed in the center of the classroom) and attempted to ensure that students sat in the same seats for each discussion. Speaker demographics varied slightly across discussions within the same classroom due to student absences and discussion styles (e.g., holding a small group discussion in which only a portion of the students were expected to speak). Descriptions of the classrooms (Table 1) reflect the maximum number of possible discussants when all students were present. Any variation between these classroom descriptions and the descriptions of the discussions within each classroom (Table 2) can be explained either by student absences or discussion style.

In most discussions, the entire class participated, although five discussions were set up as small groups and thus limited to a subset of the students present (see 6a, 6b, 6c, 7b, 7c in Table 2). Discussions were recorded by a member of the research team using a Zoom H6 six-track portable audiorecorder placed in the center of the classroom with student discussants arranged in circles around the device. In addition to creating a map that linked numerical IDs to the locations of each discussant, a researcher also kept handwritten notes to identify speakers. Transcriptions of the audio data were outsourced to a professional service (Rev.com) and were reviewed by research assistants for accuracy. In a test of four transcripts, an average of 4% words per transcript were incorrectly transcribed and required revision. In addition to aligning speaker IDs to transcribed talk, research assistants also transferred data to an excel document formatted specifically for annotation.

4. Data Annotation

The Discussion Tracker corpus includes annotations for three dimensions of student talk that researchers in classroom discourse have associated with positive educational outcomes (Howe et al., 2019; Applebee et al., 2003; Chisholm and Godley, 2011; Soter et al., 2008; Sohmer et al., 2009; Juzwik et al., 2013; Nystrand and Gamoran, 1991) (see Table 3); similar dimensions have also been annotated by NLP researchers for other types of data. Using a classroom discussion annotation scheme optimized for NLP development (Lugini et al., 2018), we annotated student talk for *argument moves* (claims, evidence and explanations) and *specificity* (low, medium and high). In addition, we developed an annotation scheme for *collaboration* that synthesized findings in both classroom discourse research (Engle and Conant, 2002; Keefer et al., 2000) as well as the computer-supported collaborative learning (CSCL) literature (Samei et al., 2014; Zhang et al., 2013).

Prior to annotation, speakers were identified using the handwritten notes taken during data collection. Cases in which speaker IDs were difficult to determine were labeled either as 'St?' if the speaker was likely a student or '?' if it was unclear whether the speaker was a student or teacher. Student talk was segmented into both turns at talk and argument discourse units. Talk at the turn level was annotated for collaboration, and talk at the argument discourse unit was annotated for argument move values (claim, evidence, explanation) and specificity level (low, medium, high). The coding instructions for all annotated dimensions are briefly

Tchr	M/F	Tchr Exp	Loc.	Grade	Course	Class size	Male	Fem.	Race White	Race Non-white
1	F	12	suburb	10	regular	29	20	9	28	1
2	M	18	suburb	11	honors	11	5	6	7	4
3	F	6	suburb	9	honors	16	11	5	8	8
4	M	12	suburb	12	AP	13	4	9	6	7
5	F	15	urban	9	regular	16	11	5	13	3
6	F	14	urban	11	AP	18	11	7	14	4
7	F	30	urban	10	regular	25	15	10	21	4
8	F	30	rural	12	AP	13	3	10	12	1
9	F	20	rural	10	regular	15	11	4	15	0
10	M	20	rural	9	honors	25	16	9	25	0

Table 1: Characteristics of 10 classroom environments. Columns from left to right: teacher, teacher's gender (male or female), years of teaching experience, school location, grade level, course type, class size, number of male students, female students, white students, non-white students.

reviewed below. More details can be found in Lugini et al. (2018), and a sample coded transcript with all annotation manuals can be found in the link provided in section 5.

Similar to argument mining systems like Nguyen (2018), our pipeline for annotating collaborative argumentation involved several steps. (1) While examining only student talk, we flagged turns that contained no substantive argumentation and were thereby deemed *non-argumentative*. Turns that included both non-argumentative and argumentative phrases were considered argumentative. In addition to talk that was inaudible or off-topic ("I have to go to the bathroom"), non-argumentative talk also included metadiscourse talk ("okay you can take a turn,"), discussion prompts ("Describe the imagery in the poem") and brief agreements ("yeah"). (2) Argumentative turns were annotated for one of four collaboration dimensions, *New Ideas*, *Agreements*, *Extensions*, and *Probes/Challenges*. (3) Turns containing collaborative argumentation were further segmented into argument discourse units, which were, (4) labeled for argument move type (*claims*, *evidence*, or *explanations*), and (5) annotated for specificity.

4.1. Collaboration

Each argumentative turn was annotated for one of four possible collaborative relationships with prior turns at talk. (1) *New Ideas*: turns that did not reference ideas in prior turns at talk. (2) *Agreements*: turns that repeated verbatim or almost verbatim the idea in a prior turn. (3) *Extensions*: turns that built on prior ideas, either the speaker's own or another student's. (4) *Probes/Challenges*: turns that questioned or disagreed with a prior idea. Also included in collaboration annotations was a reference to the prior turn with which the current turn was in a collaborative relationship (turn reference number). After coding approximately one-third of the transcripts, analyses revealed that 30% of turns had a collaborative relationship with one of the previous two turns, and 95% had a collaborative relationship with turns within the previous four turns. Thus a limit for turn references was set at no more than four annotated previous turns unless the speaker's reference to an earlier turn was explicit (e.g., a speaker said, "Going back to John's comment about authority" when John had commented 10 turns previously).

4.2. Segmentation

Prior to annotating for argumentation and specificity, argumentative turns at talk were segmented further into argument discourse units (ADUs). Similar to Ghosh et al. (2014), who segmented ADUs into either "stance" vs "rationale," annotators were instructed to divide turns at talk into interpretive vs. factual/ information-based segments of talk. For example, as seen in Table 3, Speaker 1's turn at talk was first segmented when they offer examples "throughout history" of their claim. The turn was segmented a second time when the speaker offered an interpretation of how the examples related to their claim. Annotators were not expected to get so fine-grained as Stab and Gurevych (2017b), whose ADU segmentation accounted for sub-claims and multiple premises. Thus, annotators were instructed not to segment turns into multiple claims or multiple units of evidence.

4.3. Argumentation

As in Lugini et al. (2018), annotations for argumentation were derived from classical models of argument structure (Toulmin, 1958), and were simplified to include three labels: *claim* (an arguable statement that presents an interpretation of a text or topic), *evidence* (facts or information to support a claim) and *explanation* (reasoning or justification for why the given evidence supports the claim).

4.4. Specificity

Our annotation scheme for specificity differs from Carlile et al. (2018) in that our labels were not contingent on argumentation, but rather stood independent of argumentation labels. Like Lugini et al. (2018), we defined specificity as the existence of particularity, detail, content-language (use of disciplinary terminology like "symbolism" or "irony"), and/or a chain of reasons. Argument units that included two or more of these four characteristics above were annotated as *high specificity*; argument units containing one of the characteristics were annotated as *medium specificity*; and argument units containing none of the above characteristics were annotated as *low specificity*.

Disc.	Text	Stu.	Min.	#Trns
1a	Death of Ivan Illych	27	40	208
1b	Night	28	41	134
1c	The Name	24	42	216
2a	Lgnd of Slpy Hllw	11	32	49
2b	The Mnstr's Black Veil	9	19	43
2c	Dickinson Poems	10	33	110
3a	Lord of the Flies	16	40	99
3b	To Kill A Mbird	12	35	81
3c	To Kill A Mbird	14	41	77
4a	Heart of Darkness	13	44	109
4b	Scarlett Ltr	13	45	63
4c	A Mdsmmr Night's Drm	11	39	119
5a	To Kill A Mbird	15	29	106
5b	Smthg Wckd This Wy Cms	16	35	105
5c	The Little Prince	15	38	111
6a	The Immortal Lf of H. Lcks	6	35	124
6b	The Crucible	6	38	79
6c	Into the Wild	7	33	293
7a	Of Mice and Men	25	25	192
7b	Fahr. 451	11	34	332
7c	MLK Jr.	13	38	141
8b	The Yellow Wllppr	11	39	127
8c	Antigone	13	28	248
9a	Salem Witch Trials	15	35	275
9b	The Crucible	14	25	264
9c	The Prks of Bng Wllflwer	15	38	267
10a	Bleachers	23	38	165
10b	JFK Speech	25	33	148
10c	To Kill A Mbird	25	33	188

Table 2: Overview of discussions by teacher. "Disc" = Discussion, numbers correspond to teacher in Table 1; "Text"=the titles of the texts under discussion; "Stu."= amount of students present in discussion; "Min." = length in minutes; "#trns"= number of turns per discussion

4.5. Reliability Analyses

Reliability analyses of segmentation and annotations yielded high inter-rater agreement. We calculated reliability measures for each of the following categories: 1) selecting collaborative argumentative turns at talk, 2) annotations for collaboration labels, 3) segmentation of turns into argument discourse units, 4) annotations for each argument move label and 5) specificity labels.

Recall that annotators were instructed to consider turns at talk as *non-argumentative* if they did not include argumentation. Agreement between annotators for argumentative turns versus not was based on a sample with highly skewed class distributions yielding low Kappa (0) but high raw percentage of agreement (92%).

Basing our argumentative turns segmentation metric on Habernal and Gurevych (2017), we computed agreement on 54 transcripts from a classroom corpus we collected before Discussion Tracker by comparing the segmentation of two trained annotators. The average alpha was 0.96, min was 0.72, max was 1, and standard deviation was 0.048. All data in the Discussion Tracker corpus was segmented by the same trained annotator.

A portion of the transcripts were double-annotated and yielded substantial agreement for collaboration (Cohen's Kappa, 0.74) and specificity (Quadratic Weighted Kappa, .70), and near-perfect agreement for argumentation (Cohen's Kappa, .89). Because our specificity annotation was based on an ordered scale (low, medium, high), we employed an inter-rater agreement measure (Quadratic weighted Kappa) that could account for degrees of disagreement. Because argument move annotations were not based on an ordered scale we simply used Cohen's Kappa in which disagreements were not specially weighted.

After achieving satisfactory agreement in double-coding, the remaining transcripts were single-annotated for collaboration, argumentation, and specificity. Differences between annotators were resolved through deliberation to construct gold standard annotations for public release.

4.6. Corpus Statistics

Of Discussion Tracker's 3261 student turns, 2128 were considered *argumentative* and were annotated for collaborative argumentation. As seen in Table 4, the large majority of annotated turns were labeled as either *New Ideas* (37.69%) or *Extensions* (47.70%), with *Challenges/Probes* and *Agreements* making up only a narrow portion of the corpus.

The argumentative turns at talk were further segmented into 3135 argumentative discourse units to be annotated for argument move type and specificity. The corpus is made up of mostly claims (65.30%), less than half of which were supported with evidence (24.31%), and still less were elaborated with explanations (10.40%). Specificity of argument moves was more evenly distributed with 37.93% annotated as low, 34.16% as medium, and 27.91% high.

5. Public Release

The Discussion Tracker Corpus will be freely available for research purposes, with the release coordinated with the publication of this paper. The release will include 29 separate .xlsx documents segmented for both turns at talk and

Turn	Speaker	Talk	Collaboration	Reference Turn	Argumentation	Specificity
1	St 1	My interpretation of it is that, without a middle ground, you are left with two very extreme points. Whether or not the middle ground directly centered, we have a range. We have a spectrum.[...]	New		Claim	Medium
		Throughout history, whether you go back to ancient Europe, and you look at tyrannies and dictatorships, not even ancient Europe. If you go back to the Holocaust and what Hitler was doing over in Germany [...] if you go back to Communism, as well [...]			Evidence	Medium
		Those are two extremes, and neither of them ended well, and just anarchy there. There is no order there, there is no civilized kind of society to base anything around. I think the middle ground is necessary just to create some kind of spectrum that we can go off of.			Explanation	Medium
2	St 9	I acknowledge your point, but there wasn't nobody going against anything until this happened, until this event occurred.	Challenge	1	Claim	Low
3	St 1	Does that make the way they were living right, thought?	Challenge	2	Claim	Low
4	St 9	If they were happy, I believe they were perfectly fine.	New		Claim	Low
5	St 17	My assessment of the topic at hand is, there needs to be a balance between state rights and user rights. [xx] slide, and to what extent was it off balance.	Extension	1	Claim	Medium
6	St 14	I concur with both St 19 and 17's statements. I also think that if we have society in which people are afraid to go against the core, then the rights of them are restricted. They're afraid that if they step outside the lines, then it won't end good for them, so everybody's afraid [...] they won't be accepted.	Extension	5	Claim	Medium
7	St 18	I concur with St 14.	Extension	6	Claim	Low
		Because back in the day when we have the Civil War going on, people were on different sides. People were afraid to come and say, "oh, I'm in between," because then they would be afraid that they'd be treated just like African Americans. As St.,9 said, it got to the point where they were on two different sides and they couldn't decide on something, so they said, "hey, let's fight this out, and whoever wins basically decides what action happens."			Evidence	High

Table 3: Sample transcript from discussion 6b on the play "The Crucible".

argument discourse units. Each document will contain the discussion in full, including teacher talk and non-annotated student talk for context. Additionally, transcripts will contain unique ID numbers (e.g., T127.1.Heartdark) for each turn at talk indicating the de-identified teacher (e.g., T127), the discussion (e.g., .1– referring to the first discussion from that teacher's classroom), the text (e.g., "Heartdark" for Joseph Conrad's *Heart of Darkness* and the turn number in the discussion (the final number in the ID). Directly to the right of each turn at talk will be columns containing collaboration annotations and their corresponding turn reference numbers. Talk segmented at the ADU level will include annotations for each argument move type and specificity.

The corpus is available at `https://`

	Annotation	Total Count	Percentage
Collaboration	New	802	37.69%
	Agree	37	1.74%
	Extensions	1015	47.70%
	Chall/ Probes	274	12.88%
	Total	**2128**	**100.00%**
Argumentation	Claims	2047	65.30%
	Evidence	762	24.31%
	Explanations	326	10.40%
	Total	**3135**	**100.00%**
Specificity	Low	1189	37.93%
	Medium	1071	34.16%
	High	875	27.91%
	Total	**3135**	**100.00%**

Table 4: Descriptive statistics of corpus annotations.

`discussiontracker.cs.pitt.edu`. Included in the corpus are 29 transcripts with complete annotations, a metadata file containing information for the speaker demographics and grade levels, and the coding manuals for creating the annotations.

The code for all classification experiments from Section 6. is also available via the Discussion Tracker website to provide a performance benchmark for future research that uses this corpus.

6. Case Studies Using the Corpus

Our corpus provides NLP researchers the opportunity of several uses. First, each of the three annotated dimensions of collaborative argumentation can be used individually to train a classifier for automated prediction. Second, we believe that one of the most interesting characteristics of the Discussion Tracker corpus is the fact that it provides annotations for multiple dimensions of collaborative argumentation simultaneously. It is possible, then, to analyze if and how these dimensions are related. If that is the case, it may be possible to develop more robust and accurate models for automated classification of such dimensions. Carlile et al. (2018) annotated argumentative discourse units for argument component and specificity (among other things) and were able to make use of these two aspects simultaneously to predict argument persuasiveness in written essays. Similarly, we showed in our previous study that specificity can be used to improve the performance of argument component classifiers (Lugini and Litman, 2018). We found that models trained through multi-task learning where the primary task consists of argument component classification and the secondary task consists of specificity classification almost always outperform models that only perform argument component classification. However, the corpus used in our previous study is not publicly available and therefore our previous results are not reproducible by other members of the research community.

To provide reproducible performance baselines to facilitate future classifier evaluations using the Discussion Tracker corpus, we present our experiments in learning models for individual classification tasks and then jointly learning multiple classifiers. The performance of each model was evaluated using the same ten-fold cross-validation: each fold consists of 26 transcripts as training set and 3 as test set (except for one fold where 27 transcripts are used for training and 2 for testing). We report accuracy, Cohen Kappa (quadratic-weighted for specificity and unweighted for the remaining tasks) and macro f-score as evaluation metrics. The particular folds used for the cross-validation experiments presented in this paper will be made available in the corpus release in the form of a json file containing a list of all training and test transcripts for each fold.

6.1. Learning through Individual Annotations

6.1.1. Argumentation

As a baseline for evaluating the performance of argument component classification on the Discussion Tracker corpus, we tested our previously proposed model (Lugini and Litman, 2018), which showed significantly higher performance on a previously examined set of classroom discussions compared to argument mining models developed for other types of corpora. It consists of a hybrid model which combines embeddings generated through a neural network with a set of handcrafted features.

The handcrafted features consist primarily of two sets: Speciteller feature set, derived from prior work on specificity (Li and Nenkova, 2015), and online dialogue feature set, derived from prior work on argument mining in online dialogues (Swanson et al., 2015). The Speciteller feature set contains the following features, extracted independently for each argument move: average of 100-dimensional word vectors (Turian et al., 2010) for words in the argument move, number of connectives, number of words, number of symbols, number of numbers, number of capital letters, number of stopwords normalized by argument move length, average characters per word, number of subjective and polar words (extracted using the MPQA (Wilson et al., 2009) and the General Inquirer (Stone and Hunt, 1963) dictionaries), average word familiarity (extracted using MRC Psycholinguistic Database (Wilson, 1988)), and inverse document frequency statistics (maximum, minimum). The online dialogue feature sets includes: number of pronouns, number of occurrences of words of different lengths, descriptive word-level statistics, term frequency - inverse document frequency of unigrams and bigrams (with frequency greater than 5), and part of speech tag features (unigrams, bigrams and trigrams).[1]

The neural network model is composed of a series of 3 convolutional/max-pooling layers, in which the convolutional layer consists of 16 filters of size 7. Each word in an argument move is first processed through an embedding module which uses pretrained GloVe embeddings (Pennington et al., 2014) of dimensionality 50. The argument move is then processed through the neural network to generate a fixed-size embedding. The handcrafted features are concatenated to the neural network embeddings, and the final feature vector is input to a softmax classifier to output

[1] One feature set from our prior work is not included in this study, namely wLDA (Nguyen and Litman, 2016), since it greatly increased model complexity leading to overfit.

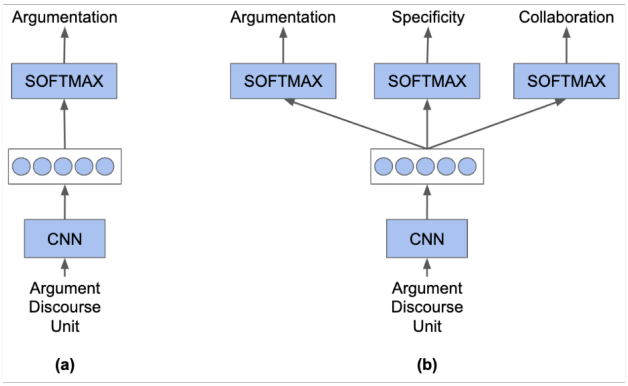

Figure 1: Neural network models used in this study: individual models (a); joint multi-task model (b).

the final prediction (see Figure 1(a)).

In our prior work (Lugini and Litman, 2018) on argument component classification for discussions, we used oversampling to alleviate the class imbalance present in argumentation labels (which is also present in the Discussion Tracker corpus). However, since our Discussion Tracker experiments also include 3 task multi-task learning, oversampling with respect to argumentation labels might have negative impact on other tasks. To address class imbalance, instead of using oversampling, we manually set class weights to impact the loss function, trying to increase the importance of the less frequent labels (evidence, explanation). The class weights were set by roughly approximating the frequency of labels in the corpus used in our prior work (Lugini and Litman, 2018): the previous corpus contained almost twice as many claims than evidence, and almost twice as many explanations as evidence, prompting us to use the weights *(claims: 1, evidence: 2, explanations: 4)*.

Table 5 shows the cross-validation results. As we can see from Row 1, the difference between accuracy and f-score indicates that the model performs differently for the three argumentation labels. The f-score for claims, evidence, and explanations are respectively 0.776, 0.565, and 0.164. While specifying class weights at training time helped, this shows that there is ample room for improvement.

6.1.2. Specificity

As we showed in our previous study (Lugini and Litman, 2017), using the off-the-shelf Speciteller tool (Li and Nenkova, 2015) for predicting sentence specificity performed poorly when applied to text-based classroom discussions. We were able to significantly improve classification results by proposing features and models explicitly developed for classroom discussions. Like our previous work on argumentation, however, the corpus is not publicly available. To provide a baseline for the Discussion Tracker corpus, we evaluated specificity prediction performance using the same model as described in Section 6.1.1. By using this model we achieved very similar performance to the model we proposed in Lugini and Litman (2017) at a fraction of the computational cost. The main difference between the two models is the use of a convolutional neural network instead of a recurrent neural network.

As we can see in Row 3 of Table 5, the small variation across the three performance metrics indicates consistent performance for all three specificity labels. Additionally, kappa and f-score show that the specificity model is much more accurate than the one for argumentation.

6.1.3. Collaboration

Since collaboration was not annotated in our previously used corpus of classroom discussions (unlike argumentation and specificity), we do not have a prior prediction model to draw upon as a baseline. We instead use Naive Bayes to model collaboration, both for model simplicity and because it is a typical baseline model for NLP tasks. The feature vector we use is bag of words (BOW) on each turn, tokenized using NLTK's word tokenizer. We filter out stop words and use tokens that occur once to approximate unknown words. We experimented with using TF-IDF weighting on bag of words and using different Naive Bayes variants. Results from the best performing configurations are reported in Table 6.

As shown in Row 1, we found that the best configuration for predicting the four collaboration labels was Multinomial Naive Bayes with TF-IDF features. Note that accuracy is much higher than both kappa and macro f-score, likely reflecting the skewed distribution among the four classes. These baseline results show the difficulty in distinguishing between all of the collaboration annotations in the Discussion Tracker corpus. We also explored a simpler binary version of our classification task (Row 2), which reduced the class skew while still making a pedagogically useful distinction. In particular, during discussions with teachers where we visualized the collaboration annotations in the corpus that came from their particular classrooms, we found that teachers were very curious about whether students were introducing new information into the discussion or building off of what was previously said. Therefore we experimented on how well a classifier could distinguish student turns labeled 'New' from the other collaboration annotations. We found that Gaussian Naive Bayes without TF-IDF features performed the best. While the results improved compared to predicting the original 4 classes, there is still room for improvement. In sum, determining the collaboration labels for student turns is difficult for our simple Naive Bayes with BOW baseline method.

Finally, although the collaboration experiment was performed using only the manually-annotated argumentative subset of student turns (as is typical of system component evaluations), an end-to-end system would in addition need to first (or jointly) automatically separate the argumentative and non-argumentative turns, before classifying the collaboration labels for the argumentative turns. Using the same approach as for predicting collaboration labels, we find that a Gaussian Naive Bayes model with TF-IDF performs the best at this task (Row 3 in Table 6).

6.2. Learning through Multiple Annotations

In this section, we describe an experiment that extends our previous multi-task learning study (Lugini and Litman, 2018), by using three rather than two tasks for the learn-

Row	Experiment	Model	Accuracy	Kappa	Macro F
1	Argument Move	Individual	0.669	0.343	0.502
2	Argument Move	Joint	0.673	0.365	0.516
3	Specificity	Individual	0.703	0.750	0.695
4	Specificity	Joint	0.706	0.751	0.698

Table 5: Neural classification results (both individual and joint models) for argument discourse unit prediction tasks.

Row	Experiment	Model	Accuracy	Kappa	Macro F
1	Collaboration (all 4 labels)	Multinomial W/TF-IDF	0.504	0.086	0.254
2	Collaboration ('New' vs Other)	Gaussian w/ BOW	0.623	0.217	0.604
3	Argumentative vs Non-Argumentative	Gaussian W/TF-IDF	0.785	0.513	0.756

Table 6: Naive Bayes classification results for turn-level prediction tasks.

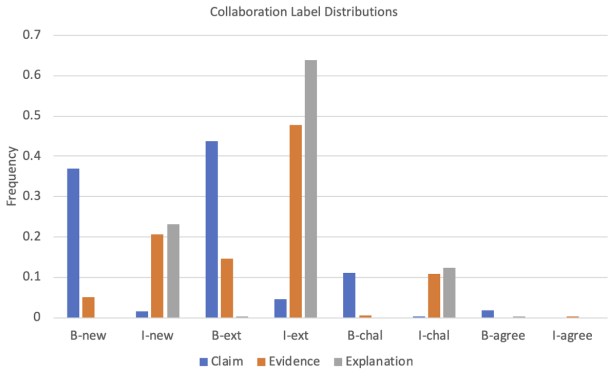

Figure 2: Distribution of collaboration labels for different argumentation categories

ing, by using the new Discussion Tracker corpus, and by providing annotation and associated benchmark results that can be replicated and extended by others in the future. More specifically, we test the following research hypotheses: by modeling argument component, specificity, and collaboration information simultaneously, we can develop a single model that outperforms individual classifiers for (H1) argument component, and (H2) specificity. These hypotheses are motivated by our observation of differences between collaboration label distributions across argumentative moves. However, given the different unit of analysis for the annotation of collaboration (turn) versus argumentation and specificity (argument discourse unit), for the multi-task learning setting the collaboration annotations have been converted to BIO format in order to have one annotation per argument move[2]. For example, we observed that the most frequent collaboration labels for claims are *B-extension* (43.9%)*, B-new* (37.0%) *and B-challenge* (11.1%). Looking at evidence, the most frequent collaboration labels are *I-extension* (47.8%)*, I-new* (20.8%) *and B-extension* (14.7%). Lastly, the most frequent collaboration codes for explanations are *I-extension* (63.9%)*, I-new* (23.2%) *and I-challenge* (12.2%). The complete label distributions are shown in Figure 2.

[2]If a turn labeled "*New*" for collaboration is segmented into two argument moves, the collaboration label is converted into "*B-New*" for the first argument move and "*I-new*" for the second.

With the goal of exploiting the potential relationships between argumentation, specificity, and collaboration, we developed a single joint model trained through multi-task learning. This model consists of the same convolutional neural network (see Section 6.1.1.) along with the same handcrafted feature set, with the difference that the final feature vector is used as input to three softmax classifiers simultaneously: one for argument component, one for specificity and one for collaboration. In this setting the representation for an argument move is entirely shared between the three tasks (see Figure 1(b)). The final loss of the model is the sum of the individual cross-entropy losses: we chose an unweighted sum so that we can understand potential relationships between the three prediction tasks; if the goal is that of maximizing performance, one of the tasks can be favored by increasing its weight in the loss function.

Table 5 shows the results of our experiments. Rows 1 and 2 relate to our hypothesis H1, and we can see an improvement in accuracy, kappa and f-score. The performance improvements achieved through the joint model, though, are not yet statistically significant. Although differences exist across argumentation labels for different collaborative moves, our joint model is not able to optimally capture them. This may be due to the low performance of the collaboration classifier: if the collaboration model cannot reliably capture collaboration information, it cannot properly inform the argumentation classifier. We believe that increasing the performance of the individual classifiers and using learned weights in the loss function will result in a more effective joint model. Rows 3 and 4 relate to our hypothesis H2. Like for argumentation, the results on specificity show the joint model outperforming the single-task one in all metrics, though the difference was not statistically significant.[3]

Although the current results show a limited gain in performance of the joint model over the individual ones, the Discussion Tracker corpus allows the research community to further analyze inter-dependencies between argumentation, specificity and collaboration, and develop more effective models to take advantage of these dependencies.

[3]Recall that while collaboration was annotated at the turn level, the joint model uses the BIO converted (ADU) representation. We thus do not investigate whether the joint ADU model improves the individual turn-level collaboration prediction.

6.3. Other Potential Corpus Uses

Going beyond classification, our corpus can also be used in conjunction with other publicly available corpora. Daxenberger et al. (2017) for example performed a qualitative analysis to understand the difference in conceptualization of claims across multiple datasets. None of the datasets analyzed, however, includes transcripts of spoken dialogues. The Discussion Tracker corpus can be used in a similar way, for example, to study the different conceptualization of argument components between spoken multi-party discussions and online multi-party dialogues or written essays. Additionally, the corpus could also support educational research, which has taken interest in classroom discourse since the 1970's (Howe and Abedin, 2013; Mercer and Dawes, 2014). Howe et al.'s (2019) recent study of 72 elementary classroom environments established statistically significant relationships between positive learning outcomes and student talk dimensions similar to the annotations we include in our corpus: participation (how much and how many students speak), elaboration (similar to our extensions), and querying (similar to our challenge/probe category). The corpus metadata can also be used to support the investigation of issues of educational equity (e.g., gender and racial) in collaborative argumentation research (Godley and Olshefski, 2019; Howe, 1997; Kelly, 2008).

7. Future Corpus Extensions

Over the next three years of the Discussion Tracker project, we will be collecting and annotating new classroom data that will more than triple the size of this first release of our corpus. In these future corpus extensions, we will also include new information in our transcripts, namely time stamps and phenomena specific to spoken data (including filled pauses like "uh"). This will allow for more investigation on the similarities and differences between spoken and written synchronous collaborative argumentation.[4]

8. Summary

By releasing the Discussion Tracker corpus, we hope to contribute to collaborative argument mining research concerned with multi-party synchronous argumentative discourse collected in authentic environments. The 29 transcriptions included in the Discussion Tracker corpus contain multi-party argumentation along with annotations for collaboration at the turn level and annotations for argument move type and specificity at the argument discourse unit level. We created performance baselines for a variety of individual classification tasks, and demonstrated the potential use of the simultaneous annotations by exploring multi-task learning as method for improving baseline performance. We believe that the Discussion Tracker corpus will be a useful resource for others, not only because it provides challenging multi-party collaborative argumentation data for future NLP research, but also because it provides multiple simultaneous annotations that can allow for a wider variety of learning approaches.

[4] Although the inclusion of audio files would contribute greatly to these endeavors and might be possible in future releases, standard IRB regulations for research on minors in authentic learning environments would require costly de-identification.

9. Acknowledgements

This work was supported by the National Science Foundation (EAGER 1842334 and 1917673) and in part by the University of Pittsburgh Center for Research Computing through the resources provided.

10. Bibliographical References

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (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.

Al Khatib, K., Wachsmuth, H., Lang, K., Herpel, J., Hagen, M., and Stein, B. (2018). Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555.

Applebee, A. N., Langer, J. A., Nystrand, M., and Gamoran, A. (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.

Ashley, K. D. and Walker, V. R. (2013). Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 176–180. ACM.

Biran, O. and Rambow, O. (2011a). Identifying justifications in written dialogs. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 162–168. IEEE.

Biran, O. and Rambow, O. (2011b). Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(04):363–381.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Cabrio, E. and Villata, S. (2014). Node: A benchmark of natural language arguments. *COMMA*, 266:449–450.

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

Chisholm, J. S. and Godley, A. J. (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.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Lan-*

guage Processing, pages 2055–2066. Association for Computational Linguistics.

Dillenbourg, P. and Hong, F. (2008). The mechanics of cscl macro scripts. *International Journal of Computer-Supported Collaborative Learning*, 3(1):5–23.

Engle, R. A. and Conant, F. R. (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.

Fischer, F., Kollar, I., Stegmann, K., and Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist*, 48(1):56–66.

Gao, Y., Zhong, Y., Preotiuc-Pietro, D., and Li, J. J. (2019). Predicting and analyzing language specificity in social media posts. *Thirty-Third AAAI Conference on Artificial Intelligence*.

Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. In *Proceedings of the first workshop on argumentation mining*, pages 39–48.

Godley, A. J. and Olshefski, C. A. (2019). Promises and limitations of applying NLP to classroom discourse analysis. Toronto, Ontario, April. Annual Meeting of the American Educational Research Association.

Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

Howe, C. and Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge journal of education*, 43(3):325–356.

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., and Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, pages 1–51.

Howe, C. (1997). Gender and classroom interaction. *SCRE PUBLICATIONS*.

Juzwik, M. M., Borsheim-Black, C., Caughlan, S., and Heintz, A. (2013). *Inspiring dialogue: Talking to learn in the English classroom*. New York.

Keefer, M. W., Zeitz, C. M., and Resnick, L. B. (2000). Judging the quality of peer-led student dialogues. *Cognition and Instruction*, 18:53–81.

Kelly, S. (2008). Race, social class, and student engagement in middle school english classrooms. *Social Science Research*, 37(2):434–448.

Ko, W.-J., Durrett, G., and Li, J. J. (2019). Domain agnostic real-valued specificity prediction. In *AAAI*.

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

Li, J. J., O'Daniel, B., Wu, Y., Zhao, W., and Nenkova, A. (2016). Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Lippi, M. and Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

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

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

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

Lugini, L., Litman, D., Godley, A., and Olshefski, C. (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.

Mercer, N. and Dawes, L. (2014). The study of talk between teachers and students, from the 1970s until the 2010s. *Oxford Review of Education*, 40(4):430–445.

Mirkin, S., Moshkowich, G., Orbach, M., Kotlerman, L., Kantor, Y., Lavee, T., Jacovi, M., Bilu, Y., Aharonov, R., and Slonim, N. (2018). Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724.

Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, Mar.

Nguyen, H. and Litman, D. J. (2016). Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics. In *FLAIRS Conference*, pages 485–490.

Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M., and Chizari, M. (2013). Facilitating argumentative knowledge construction through a transactive discussion script in cscl. *Computers & Education*, 61:59–76.

Nystrand, M. and Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, 25(3):261–290.

Park, J. and Cardie, C. (2018). A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods*

in natural language processing (EMNLP), pages 1532–1543.

Richey, C., D'Angelo, C., Alozie, N., Bratt, H., and Shriberg, E. (2016). The sri speech-based collaborative learning corpus. In *INTERSPEECH*, pages 1550–1554.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., Sun, X., Glaus, M., and Graesser, A. (2014). Domain independent assessment of dialogic properties of classroom discourse. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 233–236.

Scheuer, O., McLaren, B. M., Weinberger, A., and Niebuhr, S. (2014). Promoting critical, elaborative discussions through a collaboration script and argument diagrams. *Instructional Science*, 42(2):127–157.

Sherry, M. B. (2014). Indirect challenges and provocative paraphrases: Using cultural conflict-talk practices to promote students' dialogic participation in whole-class discussions. *Research in the Teaching of English*, pages 141–167.

Sohmer, R., Michaels, S., O'Connor, M., and Resnick, L. (2009). Guided construction of knowledge in the classroom. *Transformation of knowledge through classroom interaction*, pages 105–129.

Soter, A. O., Wilkinson, I. A., Murphy, P. K., Rudge, L., Reninger, K., and Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47(6):372–391.

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

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

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

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256. ACM.

Swanson, R., Ecker, B., and Walker, M. (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.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge University Press, Cambridge, England.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Weinberger, A. and Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & education*, 46(1):71–95.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, 20(1):6–10.

Zhang, J., Chen, M.-H., Chen, J., and Mico, T. F. (2013). Computer-supported metadiscourse to foster collective progress in wu knowledge-building communities. In *Proceedings of the International Conference of Computer-supported Collaborative Learning*.