Automated classification of collaborative problem solving interactions in simulated science tasks

Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, Alina A von Davier
Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
{mflor,syoon,jhao,lliu001,avondavier}@ets.org

Abstract

We present a novel situational task that integrates collaborative problem solving behavior with testing in a science domain. Participants engage in discourse, which is used to evaluate their collaborative skills. We present initial experiments for automatic classification of such discourse, using a novel classification schema. Considerable accuracy is achieved with just lexical features. A speech-act classifier, trained on out-of-domain data, can also be helpful.

1 Introduction

Collaborative problem solving (CPS) is a complex activity that involves an interplay between cognitive processes, such as content understanding, knowledge acquisition, action planning and execution (Greiff, 2012; von Davier and Halpin, 2013), non-cognitive processes, such as adaptability, engagement, social regulation, and affect states, such as boredom, confusion, and frustration (Baker et al., 2010; Graesser et al., 2010). Collaborative learning techniques are used extensively in educational practices, from pre-school to higher education. Collaborative activity in learning environments may take place in face-to-face interactions or via online distance-learning platforms (Prata et al., 2009).

Within the domain of educational assessment, there has been a strong recent interest in the evaluation of CPS as a social skill (Griffin et al., 2012; Liu et al., 2015; von Davier and Halpin, 2013). Such interest is informed by analysis of group interactions, which often integrate context, experience, and active engagement of learners (Hatano & Inagaki, 1991; Hmelo-Silver, Nagarajan, & Day, 2000). For example, Damon and Phelps (1989) pointed out that collaborative discussion provides a rich environment for mutual discovery, reciprocal feedback, and frequent sharing of ideas. Duschl and Osborne (2002) noted that peer collaboration provides opportunities for scientific argumentation – proposing, supporting, criticizing, evaluating, and refining ideas.

To include discursive collaboration in large-scale educational assessments, it is essential to automate the scoring and annotation process of discursive interactions. In our study, we explore an application of natural language processing techniques for annotating group discourses using a novel CPS classification framework.

The rest of this paper is structured as follows. Section 2 presents the experimental task that is used for eliciting collaborative behavior in a controlled setting. Section 3 describes the collected data. Section 4 presents the CPS classification framework and the manual annotation of data according to this framework. Machine learning experiments for automated annotation of collaborative interactions are presented in section 5.

2 Task Description

We have designed a research study to explore the relationship between CPS skills and collaboration outcomes (Hao et al., 2015). We focus on measuring collaboration skills within the domain of science. The task was structured as a computer-based simulation, in an interactive game-like environment. Such setting can provide students with opportunities to demonstrate proficiencies in complex interactive environments that traditional assessment formats cannot afford (Klopfer et al., 2009). The simulation

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 31–41,
San Diego, California, June 16, 2016.

task was modified from an existing simulation, Volcano Trialogue (Zapata-Rivera et al., 2014), and delivered over a web-based collaborative platform (see Figure 1). Task participants took the roles of assistants in a virtual seismic measurement laboratory, measuring and monitoring seismic events related to various aspects of (simulated) volcanic activity. They were given various assignments (by a script-controlled virtual scientist agent), and their performance was measured based on their responses to the assignments in the simulation task. In this task, two participants work together via text chat to complete the specific subtasks. All of the turn-by-turn conversations and responses to the questions were recorded in an activity log with time stamps. The conversations were used to measure CPS skills, while responses to the in-simulation test items were used to measure science inquiry skills.

Figure 1. Sample screenshot from the Volcano task.

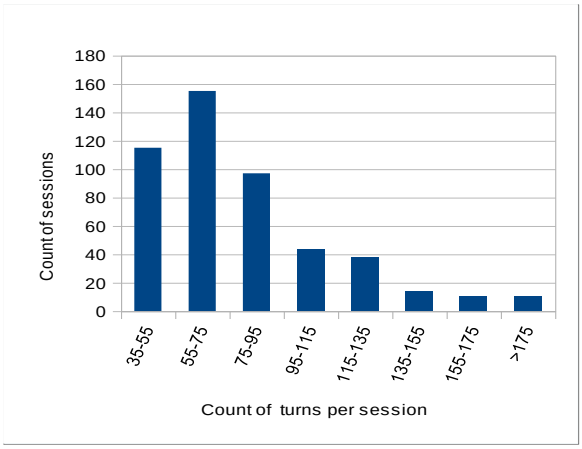

Figure 2. Binned distribution of turn counts per session

A session in this simulation task consists of multiple items/subtasks in various formats, such as multiple choice, constructed response, and conversations with virtual agents. There were also action items, such as placing seismometers on a virtual volcano map and making notes of collected seismic data. Pre-designed prompts were displayed in the system prompt area to guide participants through the sequence of subtasks in a session. To capture the evidence for the outcomes of collaboration in the simulation, a three-step response procedure was used for each item. First, each participant was prompted to respond the item individually. Next, each participant was prompted to discuss the item with the partner. Individual response could be revised at this stage and a team-response could be negotiated. Finally, a team-representative was randomly chosen to submit a team answer. The changes in the test-responses before and after the collaboration may indicate how effective the team collaboration was. In a separate paper, we describe how such changes provide insights on which CPS skills are important for better collaboration outcomes (Hao et al., submitted). In the present paper, we focus on developing automated methodologies to classify the conversations in the sessions.

3 The CPS chat data

Data was collected through the Amazon Mechanical Turk crowdsourcing data-collection platform. We recruited 1,000 participants with at least one year of college education. Participants were teamed randomly into pairs to take the collaborative science simulation task. After removing sessions with incomplete data, we had complete responses from 482 teams. Figure 2 presents a binned histogram for the amounts of turns taken in the 482 sessions, indicating the amount of dialogue that has occurred. A 'turn' consists of whatever text a participant types before pressing 'Send'. About 80% of the sessions had 35-100 turns. The chattiest session had 300 turns. Sample chat excerpts are presented in Table 2. Overall, there are 38,703 turns in our corpus. The total number of tokens is 189K (213K with punctuation). Average token-count per turn is 4.9 tokens (5.5 with punctuation).

4 CPS classification

By analyzing discourse, researchers can make sense of how students collaboratively solve problems. Observable features from the collaborative interaction, such as turn taking, sharing resources and ideas, negotiating, posing and answering questions, etc., may be used for measuring CPS skills.

There are many different ways to annotate interactions, for different purposes. Dialogue acts (DA) are sentence-level units that represent states of a dialogue, such as questions, statements, hesitations, etc. However, classification of dialogue acts differs from CPS classification (Erkens and Janssen, 2008). Whereas dialogue act coding is based on the pragmatic, linguistic features, close to utterance form, the coding of collaborative activities is based on the theoretical interpretation of the content of the utterances – the aim and function of the utterances in the collaboration process. For example, from the DA perspective, "Look at the map" and "Wait for me" are simply directives. From CPS perspective, the former may be considered "Sharing of ideas/resources", the latter – a "Regulating" expression.

Research in the field of collaboration analysis has not settled yet on a single CPS annotation scheme. Hmelo-Silver and Barrows (2008) provide a schema for characterizing collaborative knowledge building for medical students working with an expert facilitator, thus focusing on facilitation aspects. Higgins et al. (2012) present a coding scheme that is focused on the types of interactions between participants (negotiation, elaboration, independence, etc.). Asterhan and Schwarz (2009) describe dual CPS coding of discussion protocols. Kersey et al. (2009) focus on knowledge co-construction in peer interactions. Mercier et al. (2014) describe a coding scheme that focuses on leadership skills in CPS. Lu et al. (2011) describe a coding scheme of discourse moves in online discussions. Weinberger and Fischer (2006) provide a multi-dimensional coding scheme for argumentative knowledge construction in computer-supported collaborative learning.

4.1 The CPS Framework

The CPS classification schema used in the present work was developed based on review of computer-supported collaborative learning (CSCL) research findings (Barron, 2003; Dillenbourg and Traum, 2006; Griffin et al., 2012; von Davier and Halpin, 2013) and the PISA 2015 Collaborative Problem Solving Framework (OECD, 2013). Our schema (Liu et al., 2015) is comprised of 33 CPS skills grouped into four major dimensions. The full listing is presented in Table 1. The four dimensions are: sharing ideas, negotiating, regulating problem-solving activities, and maintaining communication. The first dimension – sharing ideas – considers how individuals bring divergent ideas into a collaborative conversation. For instance, participants may share their individual responses to assessment items and/or point out relevant resources that might help resolve a problem. The second dimension – negotiating ideas – is to capture evidence of the team's collaborative knowledge building and construction through negotiating with each other. The categories under this dimension include agreement/disagreement with each other, requesting clarification, elaborating/rephrasing other's ideas, identifying gaps, revising one's own idea. The third dimension – regulating problem-solving activities – focuses on the collaborative regulation aspect of the team discourse. This dimension includes such categories as identifying goals, evaluating teamwork, and checking understanding. The last dimension – maintaining a positive communication atmosphere – is to capture social communications beyond the task-specific interactions.

4.2 Human coding of CPS classes

Two human annotators were trained to annotate the chats. Training involved overview of definitions and coding examples for each of the 33 categories of CPS skills. After training, annotators independently coded discourse data from the chat protocols. Seventy seven sessions out of 482 (16%) were coded by both annotators, all other sessions were coded by the same single annotator (H1).

The unit of analysis was each turn of a conversation, i.e. each turn received a label drawn from the 33 categories. Due to complexity of the collaborative process, one turn of chat may have more than one function that can be mapped in the CPS framework. Therefore, an annotator was allowed to assign up to two labels to each turn. A primary label reflects what the annotator considered as the major function in a given turn, and a secondary label reflects an additional, less central function.

Table 2 presents a sample of this annotation. The first column marks speaker-ID, the second column

CPS skills	Student performance (categories)
Sharing ideas	1. Student gives task-relevant information (e.g., individual response) to the teammate. 2. Student points out a resource to retrieve task-relevant information. 3. Student responds to the teammate's request for task-relevant information.
Negotiating ideas	4. Student expresses agreement with the teammates. 5. Student expresses disagreement with teammates. 6. Student expresses uncertainty of agree or disagree. 7. Student asks the teammate to repeat a statement. 8. Student asks the teammate to clarify a statement. 9. Student rephrases/complete the teammate's statement. 10. Student identifies a conflict in his or her own idea and the teammate's idea. 11. Student uses relevant evidence to point out some gap in the teammate's statement. 12. Student elaborates on his or her own statement. 13. Student changes his or her own idea after listening to the teammate's reasoning
Regulating problem solving	14. Student identifies the goal of the conversation. 15. Student suggests the next step for the group to take. 16. Student expresses confusion/frustration or lack of understanding. 17. Student expresses progress in understanding. 18. Student reflects on what the group did. 19. Student expresses what is missing in the teamwork to solve the problem. 20. Student checks on understanding. 21. Student evaluates whether certain group contribution is useful or not for the problem solving. 22. Student shows satisfaction with the group performance. 23. Student points out some gap in a group decision. 24. Student identifies a problem in problem solving.
Maintaining communication	25. Student responds to the teammate's question (using texts and text symbols). 26. Student manages to make the conversation alive (using texts and text symbols, using socially appropriate language). 27. Student waits for the teammate to finish his/her statement before taking turns. 28. Student uses socially appropriate language (e.g., greeting). 29. Student offers help. 30. Student apologizes for unintentional interruption. 31. Student rejects the teammate's suggestions without an accountable reason. 32. Student inputs something that does not make sense. 33. Student shows understanding of the teammate's frustration.

Table 1: CPS Framework coding rubric of collaborative problem solving interactions skills

presents the chat text. The fourth column presents the primary classification code assigned by annotator H1. The third column indicates the general dimension for primary label. The fifth column shows the secondary labels given by annotator H1. For example, in the last row of the table, note that the response of participant s8 is classified as primary=category#11 (point out a gap in statement) and secondary=category#2 (suggest information resource).

As the annotation focused on marking only the most prominent CPS functions, a secondary label was used only if the annotator considered that the additional function was prominent enough. Our first annotator assigned a secondary tag (in addition to a primary tag) to 13,404 chat-turns (34% of all cases), while the second annotator used a secondary tag in only 33 cases. Thus, we disregard the secondary tag of the second annotator and compute inter-rater

ID	Chat text	Dimension	P	S
s1	hi there i'm george!	maintaining communication	28	
s2	good morning, I'm j	maintaining communication	28	
s1	i think the answer is b	sharing ideas	1	
s2	magma approaching crater?	sharing ideas	1	20
s1	i remember the video saying high frequency waves resulted from rocks cracking	sharing ideas	2	12
s2	i just reviewed it and you are correct	negotiating ideas	12	4
s2	c	sharing ideas	1	
s1	i chose c as well	sharing ideas	1	
(some turns omitted here)				
s2	?	regulating	16	
s1	we can keep the first two notes we seemed to have similar answers	regulating	15	
Example from another session:				
s9	I thought it started low?	sharing ideas	1	
s8	nope you can look at the volcanic seismic events in the bottom left corner	negotiating ideas	11	2

Table 2: Sample excerpts from some chat sessions, with CPS classifications by annotator H1. P=primary label, S-secondary label. Labels are explained in Table 1.

agreement as follows. The 77 sessions that were processed by both annotators had 6,079 turns. Over those turns, the two annotators agreed in their primary tags in 3,818 cases (62.8% agreement, Cohen's kappa 0.56). We also considered a different criterion, when second annotator's primary tag agreed with either primary or secondary tag given by the first annotator. In this approach, agreement is 72%s and Cohen's kappa is 0.62. According to Landis and Koch (1977) scale, those levels of agreement are somewhere between moderate (kappa 0.41-0.61) and substantial (kappa 0.61-0.80).

Figure 3 presents the distribution of primary and secondary CPS labels assigned by annotator H1 to the whole set of 38,703 turns. The distribution is very uneven. Two categories were never used (#26 and #27).

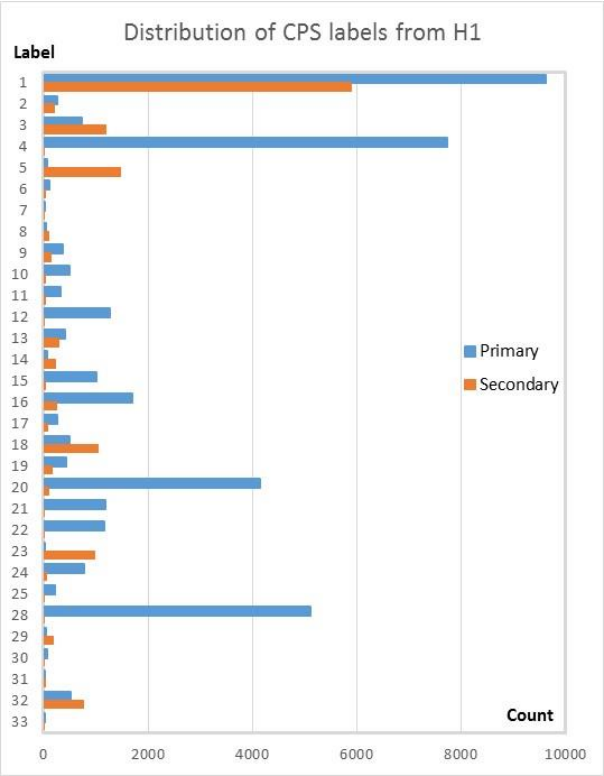

Figure 3. Histogram of primary and secondary CPS labels assigned by annotator H1 for the set of 38,703 chat turns.

It is not uncommon to see uneven distribution of categories in collaborative discourse (Chin and Osborne, 2010; Schellens and Valcke, 2005; Lipponen et al. 2003). Given that the task prompted students to share individual responses, it is also not surprising to see categories #1 (give info to partner) and #4 (expresses agreement) as the most frequent codes. Social factors may also be at play. For instance, people often tend to be polite and respectful and express disagreement indirectly. Instead of saying "*no, I disagree*", very likely a person would say "*my answer is …*" or "*I think it's … because*", and such responses are not coded as expressed disagreement, but rather as sharing or negotiating. This may explain why explicit agreements are five times more frequent than explicit disagreements in our data, the latter also mostly coded as secondary label.

5 Automation of CPS classification

Analysis of protocols and logs of communication is an important research technique for investigating collaborative processes. Since such analysis is very

time consuming, researchers have turned to automating such analyses by utilizing computational linguistic methods. Discourse Processing is a well-established area in computational linguistics, and there is much research on automatically recognizing and tagging dialogue acts, for spoken or transcribed data (Webb & Liu, 2008; Webb et al., 2005; Serafin & Di Eugenio, 2004; Stolcke et al., 2000; Samuel et al., 1999). De Felice and Deane (2012) describe identifying speech acts in e-mails, as part of a scoring system for educational assessment. Similarly, researchers have developed computational linguistic approaches in analysis of collaboration protocols. Law et al. (2007) presented a mixed system, where manual coding is augmented with automated suggestions, derived from keyword and phrase matching. Erkens and Janssen (2008) describe a rule-based approach for automatic coding of dialogue acts in collaboration protocols. Erkens and Janssen have also stressed the difference between dialogue acts, which are closer to the linguistic form of interaction, and classes of collaborative utterances, which are more context specific and depend on respective theoretical frameworks of collaboration processes. Rosé et al. (2008) and Dönmez et al. (2005) describe a machine learning approach to code collaboration protocols according to the classification system of Weinberger and Fischer (2006).

Here we describe machine learning experiments towards automated coding of collaboration protocols according to the novel CPS framework. In our experiments we attempt to learn directly the 31 categories of the 33 defined in the framework. We chose a multinomial Naive-Bayes and HMM approaches, as starting points to explore assigning CPS tags to chat-turns in our data.

As a pre-processing step, all texts were automatically spell-corrected using a contextually-aware spell checker (Flor, 2012). Slang words and expressions were normalized (e.g. {*ya, yea, yeah, yiss, yisss, yep, yay, yaaaay, yupp*} → *yes*), using a dictionary of slang terms (this functionality is provided in the spell-checker). All texts were then tokenized and converted to lower case.

Following the manual annotation, our goal is to provide a single class-label to each chat turn, which may consist of one or more sentences. All machine learning experiments reported here use five-fold cross-validation with 4:1 train:test ratio. For this purpose the 482 collaboration sessions (described in section 3) were partitioned (once) into five groups: two groups of 97 sessions and three groups of 96 sessions each). This also resulted in unequal (but approximately similar) amount of turns in each fold (7541 turns in the smallest fold, 8032 in the largest).

Experiment 1. In our first experiment we train a Naive-Bayes classifier on just the primary labels from human annotator H1. We do not filter features, but rather use all available tokens. We use lexical features (word and punctuation tokens) in several configurations – unigrams, bigrams and trigrams. Performance results (micro-averaged across test-folds) are shown in Table 3. As a baseline, we always predict the most frequent category (CPS category#1), an approach that achieves 24.9% accuracy. The best result, 59.2% classification accuracy, is achieved by using just unigram features (single words and punctuation tokens). It clearly outperforms the baseline by a large margin. Bigrams and trigrams are not useful, their use actually decreases classification accuracy, as compared to using just unigrams.

We also experimented with ignoring the punctuation. A Naïve-Bayes classifier trained on lexical unigrams, but without punctuation, achieves accuracy of 55.5%. This is lower than the 59.2% achieved when punctuation is used. It demonstrates that punctuation is clearly useful for classification, which is consistent with similar results in a different domain (De Felice and Deane, 2012).

Experiment 2. Since collaborative interactions in the Volcano task are clearly dialogic, it is reasonable to expect that a CPS label for a given chat-turn may probabilistically depend on the labels of previous turns (as is often the case for dialogue-acts, e.g. Stolcke et al., 2000). Thus, we explore the use of Hidden Markov Model (HMM) classifier in this case (following the approach of Stolcke et al., 2000). We explored a range of parameter settings, using n-best labels from 4 to 7 (for a single chat-turn) and look-back history of one or two turns. Looking back is restricted because the dialogue is usually localized, just a few turns focusing on the specific subtask that participants were working on. Results are presented in Table 3. HMM modeling is clearly not effective in this task, as its results are much lower than those from a Naïve-Bayes classifier. Notably, this result is not without precedent. Serafin & Di Eugenio (2004), working on dialogue

act modeling, found that using dialogue history worsens rather than improves performance.

A per-CPS-category performance comparison was conducted between the Naïve-Bayes unigrams-based classifier and the HMM classifier (with n-best=4, lookback=1). The HMM classifier performs worse than NB classifier on all categories, except CPS category #4 (student expresses agreement with the teammates), where HMM is better than NB by 7.34%. This suggests that selective integration of contextual information might be useful.

Method	Acc.%	Kappa
Baseline (most frequent class)	24.9	0.01
Experiment 1		
NB with lexical unigrams	59.2	0.52
NB with unigrams+bigrams	58.5	0.51
NB with 1-,2- & 3-grams	58.2	0.51
NB, unigrams, no punctuation	55.6	0.48
Experiment 2		
HMM, n-best=4, lookback=1	52.5	0.42
HMM, n-best=7, lookback=1	48.1	0.36
HMM, n-best=4, lookback=2	46.4	0.34
HMM, n-best=7, lookback=2	41.7	0.28
Experiment 3		
NB on CPS data, with probabilistic dialog-act tagging trained on out-of-domain data	44.6	0.30
Same as above, +lexical unigrams from CPS data	60.3	0.54

Table 3: Evaluation results (accuracy and Cohen's kappa) for machine learning classification experiments with 31 CPS Framework tag categories. All results were micro-averaged from five cross-validation test folds (N=38,703 chat turns). NB=Naïve-Bayes, HMM=Hidden Markov Model.

Experiment 3. In this experiment we investigated whether automatic dialogue-act detection, trained on out-of-domain data, can be beneficial for CPS classification. Webb & Liu (2008) have demonstrated that using out-of-domain data can contribute to more robust classification of speech acts for in-domain data. Here, we extend this idea even further – use out-of-domain training data and a different classification schema.

In a separate study, we developed an automated speech-act classifier using the ICSI Meeting Recorder Dialogue Act (MRDA) corpus (Shriberg et al., 2004). The dialogue act of each sentence was annotated using MRDA annotation scheme developed for multiparty natural human-human dialogue. It defined a set of primitive communicative actions. In total, it includes 50 tags: 11 general tags and 39 specific tags. General tags include speech act types such as statements and yes/no questions and also make distinctions among syntactically different forms. Specific tags describe the purpose of the utterance, e.g., whether the speaker is making a suggestion or responding positively to a question.

The automated speech-act classifier was trained using a set of linguistic features described in Webb and Liu (2008), including sentence length, sentence initial and final word/POS n-grams, and presence/absence of cue phrases. A Maximum Entropy model-based classifier was trained on randomly selected 40 meetings and tested on the remained 24 meetings. The kappa between the system and human annotator was 0.71 for general tag and 0.61 for specific tag. The inter-rater agreement based on the subset of data was 0.80 for general tag and 0.71 for specific tag.

Notably, the MRDA data – conversations among computer scientists – is different from our CPS data, and the tag-set of dialogue acts is different from the CPS Framework tag-set. For our experiment, we used the out-of-domain-trained speech-act classifier to process CPS chat data and recorded the predicted probabilities of each speech act. Since CPS data was not annotated for speech acts, we do not know how accurate that classifier is. We just took the assigned speech-act tags and used them as probabilistic features in training our Naïve-Bayes CPS classifier, as follows. In training a standard Naïve-Bayes text classifier, each token (e.g. word) is a feature and its basic weight (count) is 1. Thus, the standard approach is to maximize:

$$\underset{c}{\operatorname{argmax}} \log(P(C)) + \sum_i \log\ \mathrm{P}(\mathrm{w_i}|\mathrm{C})$$

We use the predicted speech-act tags as special features, and their probabilities as feature weights, using the following formula:

$$\underset{c}{\operatorname{argmax}} \log(P(C)) + \sum_i \log\left[\mathrm{P}(\mathrm{f_i}) \times \mathrm{P}(\mathrm{f_i}|\mathrm{C})\right]$$

where $P(f_i)$ is the probability of speech act f_i in the current chat-turn and $P(f_i|C)$ is the conditional probability of speech act f_i given a specific CPS tag (this part is learned in the training).

A Naïve-Bayes CPS classifier, trained with probabilistic speech-act features, achieves accuracy of 44.6% in assigning CPS tags, which is substantially better than our baseline of 24.9%. We then trained a Naïve-Bayes classifier that integrates both lexical features (unigrams) from CPS training data and probabilistic speech-act features, using the formula:

$$\underset{c}{\operatorname{argmax}} \log(P(C)) + \sum_j \log\, P(w_j|C) + \sum_i \log\left[P(f_i) \times P(f_i|C)\right]$$

where w_j are lexical features (tokens). This approach makes the additional naïve assumption that speech acts as features are independent of the words. This classifier achieves 60.3% accuracy in CPS classification, 1% more than lexical-unigrams Naïve-Bayes (significant difference, $p<0.000001$, McNemar's test for correlated proportions).

We conducted additional investigations, to look how the classifiers perform for each individual CPS category. Table 4 presents the accuracy of the Naïve-Bayes unigram-based classifier on each CPS category. One obvious conclusion is that the larger is the count of a given category, the higher is the classifier accuracy. In fact, the Pearson correlation between tag count and accuracy is 0.701. While this might be expected, there are also some examples that do not follow this trend. The classifier achieves only 9.76% accuracy on label #12 (student elaborates on own statement), although it is a rather frequent category. A possible reason for this might be that elaboration statements are highly heterogeneous in form and lexical content, their status as 'elaborations' requires some abstraction and semantic inference. Another example is label #3 (student responds to the teammate's request for task-relevant information), with only 2.31% classification accuracy. This looks like one of the cases that could benefit from considering content from previous chat-turns, although not always from the immediately preceding turn.

Detailed analysis of classifier performance on each CPS category provides some interesting findings. In experiment 3 we have found that adding probabilistic dialogue act detection as features to a lexical Naïve-Bayes classifier improves overall accuracy by just 1%. However, a detailed view reveals additional information (see last column in Table 4). For some CPS categories, adding dialogue acts considerably improves classifier accuracy: 8% for category #20 and 10% for category #22. This is not unexpected – CPS category #20 (student checks on understanding) directly corresponds to dialogue-act category 'understanding-Check'. Another case is CPS category #17 (student expresses progress in understanding), which corresponds rather directly to dialogue-act "acknowledgement". For several other CPS categories, detection of dialogue-acts was not helpful for CPS classification. In future work, we will consider how to use dialogue-act detection selectively in CPS classification.

While the results from our experiments are encouraging, higher levels of accuracy are needed for

CPS tag	Total Count	Accuracy NB	Accuracy NB DA	Change
1	9628	66.99	67.60	0.61
4	7736	83.85	82.38	*-1.47*
28	5119	67.49	66.52	-0.98
20	4136	59.99	68.01	**8.03**
16	1704	44.72	46.19	1.47
12	1260	9.76	11.43	1.67
21	1189	44.66	45.33	0.67
22	1169	49.44	59.54	**10.09**
15	1021	44.27	44.47	0.20
24	768	50.39	51.30	0.91
3	735	2.31	2.31	0.00
32	512	60.55	58.98	-1.56
18	506	29.05	28.46	-0.59
10	498	6.83	7.43	0.60
19	435	54.02	54.71	0.69
13	408	30.39	29.41	-0.98
9	359	11.98	12.81	0.84
11	333	16.52	17.12	0.60
2	261	45.21	42.53	*-2.68*
17	252	19.05	25.40	**6.35**
25	228	11.84	7.46	-4.39
6	115	13.91	13.04	-0.87
5	66	1.52	6.06	**4.55**
30	65	16.92	23.08	**6.15**
14	57	10.53	7.02	*-3.51*
29	46	8.70	4.35	*-4.35*
8	41	7.32	0.00	*-7.32*
33	26	0.00	0.00	0.00
23	16	0.00	0.00	0.00
31	11	0.00	0.00	0.00
7	3	0.00	0.00	0.00

Table 4: CPS categories, with counts (primary label by annotator H1), and average automated classification accuracy. (NB)=Naïve-Bayes with unigram features, (NB DA)= Naïve-Bayes with unigram features and Dialogue Act features.

using automated CPS classifiers in operational settings. Beigman-Klebanov and Beigman (2014) and Jamison and Gurevych (2015) have suggested that, in supervised machine learning, the presence of difficult items in the training sets is detrimental to learning performance and that performance can be improved if systems are trained on only easy data. They define 'easy' as less controversial in human annotations. We explore this aspect using the Naïve-Bayes classifier with unigram lexical features.

Experiment 4. Our annotator H1 used secondary labels when a chat-turn had two prominent functions (rather than just one). Such cases can be considered ambiguous and more difficult than cases that have only one prominent CPS function. In this experiment we filtered such cases out (reduction of about 34%, to N=25,299), from either training, testing or both. Results are shown in Table 5.

Experiment 5. Here we consider as 'easy' only those 3,818 cases where two human annotators agreed on the primary label. We train a new set of classifiers, using the same five-fold cross-validation splits, but filter out from training all cases that lack explicit consensus. Micro-averaged performance for this type of classifier is compared to the classifier that used unfiltered training data. Results are presented in Table 6.

		Testing	
		Unfiltered data	Cases without secondary tag
Training	Unfiltered data	59.2% *k=0.52*	69.0% *k=0.62*
	Cases without secondary tag	57.8% *k=0.49*	70.9% *k=0.63*

Table 5: Classifier accuracy when trained and tested with/without cases that have secondary CPS labels.

		Testing	
		Unfiltered data	Consensus cases
Training	Unfiltered data	59.2% *k=0.52*	73.1% *k=0.67*
	Consensus cases	55.3% *k=0.46*	74.4% *k=0.68*

Table 6: Classifier accuracy when trained and tested with all or just human-rater-consensus cases.

In both experiments 4 and 5 we see that when a classifier is trained on 'easy' data and tested on all data, performance degrades (relative to a classifier that was trained on all data), but degradation is very moderate. In experiment 4, training data was reduced by 30%, but degradation of accuracy was just 1.4%. In experiment 5 training data was reduced by 90%, but degradation of accuracy was just 3.9%. On the other hand, when tested on only easy data, classifiers that were trained on easy data outperform the classifiers that were trained on unfiltered data, but only by a very small margin (1-2%).

6 Conclusions

In this paper we presented a novel task that integrates collaborative problem solving behavior with testing in a science domain. For integration in educational assessment, the task would benefit from automated scoring of CPS discourse. We used a complex CPS coding scheme with four major dimensions and 33 classes. In our initial exploration, we sought to obtain a single CPS category-label for each turn in chat dialogues. Our results indicate that considerable accuracy (59.2%) can be achieved by using a simple Naïve-Bayes classifier with unigram lexical features. This result approaches human inter-rater agreement (62.8%).

For future research we consider pursuing several complementary lines of work. One direction is to use more sophisticated machine-learning approaches, such as CRF and SVM, and additional features, such as part-of-speech tags and timing of chat turns. Another direction is to explore the reasons for disagreement in human annotations. Given the complex nature of collaborative discourse, it is usual that some discourse turns carry more than one function mapped in the CPS framework. Thus, another line of exploration is to train a system to decide in which cases it may suggest more than one tag to a given chat turn, i.e. consider multi-label classification of CPS data. Finally, it might be fruitful to provide a bridge between the high-level functionally-defined CPS categories and more linguistically-oriented dialogue acts. We have shown that using a Dialogue Acts classifier, trained on out-of-domain data, can be useful for classifying CPS skills. We will explore whether an explicit mapping between dialogue acts and CPS categories may contribute to better CPS classifications.

References

Asterhan, C.S.C., and B.B. Schwarz. 2009. Argumentation and Explanation in Conceptual Change: Indications from Protocol Analyses of Peer-to-Peer Dialog. *Cognitive Science*, 33, 374–400.

Baker, R., D'Mello, S.K., Rodrigo, M.M.T. and A.C. Graesser. 2010. Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68, 4, 223-241.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 12(3):307-359.

Beigman Klebanov, B. and E. Beigman. 2014. Difficult Cases: From Data to Learning, and Back. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 390–396.

Boyer, K.E., Ha, E.Y., Phillips, R., Wallis, M.D., Vouk, M.A. and J.C. Lester. 2010. Dialogue act modeling in a complex task-oriented domain. In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 297-305. Association for Computational Linguistics.

Chin, C., and J. Osborne. 2010. Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching*, 47(7), 883-908.

Damon, W., and E. Phelps. 1989. Critical distinction among three approaches to peer education. *International Journal of Educational Research*, 13, 9-19.

De Felice R. and P. Deane. 2012. Identifying Speech Acts in E-Mails: Toward Automated Scoring of the TOEIC® E-Mail Task. ETS Research Report RR-12-16. Educational Testing Service, Princeton, NJ.

Dillenbourg, P. and D. Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121-151.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., and F. Fischer. 2005. Supporting CSCL with automatic corpus analysis technology. In T. Koschmann, D.D. Suthers, and T.-W. Chan (Eds.), Proceedings of the 2005 conference on Computer support for collaborative learning: The next 10 years! Mahwah, NJ: Lawrence Erlbaum Associates, pages 125–134.

Duschl, R., and J. Osborne. 2002. Supporting and promoting argumentation discourse. *Studies in Science Education*, 38, 39-72.

Erkens, G. and J. Janssen. 2008. Automatic coding of dialogue acts in collaboration protocols. *Computer-Supported Collaborative Learning*, 3:447–470.

Flor, M. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues* (TAL), 53:3, 61-99

Graesser, A.C., D'Mello, S.K., Craig, S.D., Witherspoon, A., Sullins, J., McDaniel, B. and B. Gholson. 2008. The relationship between affective states and dialog patterns during interactions with AutoTutor. *Journal of Interactive Learning Research*, 19, 2,293-312.

Greiff. S. 2012. From interactive to collaborative problem solving: Current issues in the Programme for International Student Assessment. *Review of Psychology*, 19, 2, 111–121.

Griffin, P., Care, E., and B. McGaw. 2012. The changing role of education and schools. In P. Griffin, B. McGaw, and E. Care (Eds.), *Assessment and teaching 21st century skills*, pages 1-15. Heidelberg, Germany: Springer.

Hao, J., Liu, L., von Davier, A., and P. Kyllonen. 2015. Assessing collaborative problem solving with simulation based tasks. In Proceeding of the 11th international conference on computer supported collaborative learning, Gothenburg, Sweden.

Hao, J., Liu, L., von Davier, A., Kyllonen, P. and C. Kitchen. (Submitted). Collaborative problem solving skills versus collaboration outcomes: findings from statistical analysis and data mining, The 9th International Conference on Educational Data Mining

Hatano, G., and K. Inagaki. 1991. Sharing cognition through collective comprehension activity. In L.B. Resnick, J.M. Levine, & S.D. Teasley (Eds.), *Perspectives on socially shared cognition*, pages 331-348, Washington, DC: American Psychological Association.

Higgins, S., Mercier, E., Burd, L. and A. Joyce-Gibbons. 2012. Multi-touch tables and collaborative learning. *British Journal of Educational Technology*, 43(6), 1041-1054

Hmelo-Silver, C.E., and H.S. Barrows. 2008. Facilitating collaborative knowledge building. *Cognition and Instruction*, 26, 48-94.

Hmelo-Silver, C.E., Nagarajan, A., and R.S. Day. 2000. Effects of high and low prior knowledge on construction of a joint problem space. *Journal of Experimental Education*, 69, 36-57.

Jamison, E.K. and I. Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 291–297.

Kersey, C., Di Eugenio, B., Jordan, P., and S. Katz. 2009. Knowledge co-construction and initiative in peer learning interactions. In Proceedings of AIED 2009, the 14th International Conference on Artificial Intelligence in Education, pages 325-332. Brighton, UK.

Klopfer, E., Osterweil, S., Groff, J. and J. Haas. 2009. *Using the technology of today, in the classroom today*. The Education Arcade, MIT.

Law, N., Yuen, J., Huang, R., Li, Y., and N. Pan. 2007. A learnable content & participation analysis toolkit for assessing CSCL learning outcomes and processes. In C. A. Chinn, G. Erkens, and S. Puntambekar (Eds.), Mice, minds, and society: *The Computer Supported Collaborative Learning* (CSCL) Conference 2007 (Vol. 8, pp. 408–417). New Brunswick, NJ: International Society of the Learning Sciences.

Landis J.R. and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

Lipponen, L., Rahikainen, M., Lallimo, J., and K. Hakkarainen. (2003). Patterns of participation and discourse in elementary students' computer-supported collaborative learning. *Learning and instruction*, 13(5), 487-509.

Liu, L., Hao, J., von Davier, A.A., Kyllonen, P. and D. Zapata-Rivera. 2015. A tough nut to crack: Measuring collaborative problem solving. Handbook of Research on Technology Tools for Real-World Skill Development, pages 344-359.

Lu, J., Chiu, M. M., and N. Law. 2011. Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior*, 27, 946-955.

Mercier, E.M., Higgins, S. and L. da Costa. 2014. Different leaders: Emergent organizational and intellectual leadership in collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning*, 9(4), 397-432.

OECD. 2013. PISA 2015 draft collaborative problem solving assessment framework. OECD Publishing, Organization for Economic Co-operation and Development. http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Prata, D.N., Baker, R., Costa, E., Rosé, C.P., Cui, Y. and A.M.J.B. de Carvalho. 2009. Detecting and Understanding the Impact of Cognitive and Interpersonal Conflict in Computer Supported Collaborative Learning Environments. In Proceedings of the 2nd International Conference on Educational Data Mining, 131-140.

Rosé, C.P., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A. and F. Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer Supported Collaborative Learning*, 3, 237–271.

Samuel, K., Carberry, S., & K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. In Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada.

Schellens, T., and M. Valcke. 2005. Collaborative learning in asynchronous discussion groups: What about the impact on cognitive processing?. *Computers in Human Behavior*, 21(6), 957-975.

Serafin, R., and B. Di Eugenio. 2004. FLSA: Extending Latent Semantic Analysis with features for dialogue act classification s. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. and H. Carvey, 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at NAACL-HLT 2004 conference.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3), pages 339-373.

Von Davier, A.A., and P.F. Halpin. 2013. Collaborative Problem Solving and the Assessment of Cognitive Skills: Psychometric Considerations. ETS Research Report RR-13-41. Educational Testing Service, Princeton, NJ.

Webb, N., Hepple, M., and Y. Wilks. 2005. Dialogue Act Classification Based on Intra-Utterance Features. In Proceedings of the AAAI Workshop on Spoken Language Understanding.

Webb, N., and T. Liu, 2008. Investigating the portability of corpus derived cue phrases for dialogue act classification. In Proceedings of the 22nd International Conference on Computational Linguistics, Vol. 1, pages 977–984, Association for Computational Linguistics.

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

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., and I.R. Katz. 2014. Assessing science inquiry skills using trialogues. In S. Trausan-Matu, K. Boyer, M. Crosby and K. Panourgia (Eds.), *Intelligent Tutoring Systems* (Vol. 8474, pp. 625-626), Springer International Publishing.