Capturing Difficulty Expressions in Student Online Q&A Discussions

Jaebong Yoo
Samsung Electronics
Suwon, South Korea
jaebong.yoo@samsung.com

Jihie Kim
KT and USC / Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA USA
jihie@isi.edu

Abstract

We introduce a new application of online dialogue analysis: supporting pedagogical assessment of online Q&A discussions. Extending the existing speech act framework, we capture common emotional expressions that often appear in student discussions, such as frustration and degree of certainty, and present a viable approach for the classification. We demonstrate how such dialogue information can be used in analyzing student discussions and identifying difficulties. In particular, the difficulty expressions are aligned to discussion patterns and student performance. We found that frustration occurs more frequently in longer discussions. The students who frequently express frustration tend to get lower grades than others. On the other hand, frequency of high certainty expressions is positively correlated with the performance. We expect such dialogue analyses can become a powerful assessment tool for instructors and education researchers.

Introduction

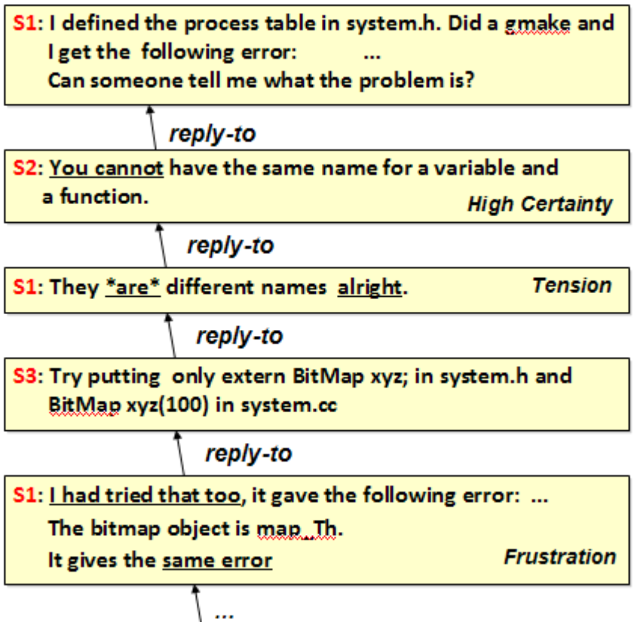
Online asynchronous discussions play an increasingly important role in sharing and exchanging knowledge in diverse fields including science, politics, health, and education. Recent studies have pointed to the discussion boards, often in the form of Q&A threads, as a promising strategy for collaboration and knowledge building (Scandamalia and Bereiter, 1994). Engagement in online discussions is also an important part in distance education and in increasingly adopted Massive Open Online Courses (MOOC). However, as such courses become more successful, their enrollments increase, and the heavier online interaction places considerable burdens on instructors.

Thus, the ultimate success of online education is constrained by limited instructor time and availability. It is probably not feasible or pedagogically appropriate to automate completely the assessment of discussion contributions. However, if we can find a way to semi-automate some part, then instructor time can be allocated to the particular students or discussion cases that truly require in-depth human monitoring and assessment.

Existing NLP work on online discussions can provide useful components to this application. Speech act analyses reveal roles that individual messages or participants play (e.g., Jeong et al., 2009). Analyses on argumentation styles (Cabrio and Villata 2012), discussion summarization (Hovy, 2006), and topic mining (Diao et al, 2012) can present an overview of how discussions go. Evaluation of user expertise or contribution quality is also useful for evaluating participant contributions (Chen et al., 2011). Current work on pedagogical analyses of multi-party online discussions covers argumentation styles (McLaren et al., 2010; Jeong 2009) or linguistic features (Morgan et al., 2011). Dialog roles in exchanging questions and answers have been characterized by (Ravi and Kim, 2007).

Affect information can support additional types of pedagogical assessment, such as problems that the participants raised and difficulties that they encounter in solving them. For example, students may express their frustration when the answer is not easily found or suggestions do not work. The degree of certainty expressed by a discussant may indicate strengths or weaknesses of the person for the given assignment topic. Figure 1 shows an example discussion from a computer science course that includes expressions of certainty and frustration. Although affect is one of the key topics in tutorial dialogue research, there is limited work on emotional expressions in Q&A discussions including how affect relates to discussion development or student performance.

In this paper, we explore a new application of online discussion analysis: how emotional expressions can be captured and used for pedagogical assessment of online discussions. Our discussion analysis framework builds on the existing Speech Act model that has been effectively


S1: I defined the process table in system.h. Did a gmake and I get the following error: ... Can someone tell me what the problem is?

S2: You cannot have the same name for a variable and a function.

S1: They *are* different names alright.

S3: Try putting only extern BitMap xyz; in system.h and BitMap xyz(100) in system.cc

S1: I had tried that too, it gave the following error: ... The bitmap object is map_Th. It gives the same error

Figure 1: An Example Discussion Thread with Emotional / Attitude Expressions.

used in representing roles of dialogue turns (Searle 1969; Cohen et al., 2004).

We define a set of emotional roles (such as frustration and degree of certainty) and informational roles (such as information seeking and information providing) of individual messages that are common in student Q&A discussions. We make use of such dialogue information in analyzing how student questions are answered and identifying difficulties that participants encounter in the process. We relate emotional and informational roles with discussion thread development and resulting performance.

The current results indicate that emotional as well as informational roles are useful in identifying student weaknesses. For example, frustration is associated with longer discussions, and frequency of frustration expressions by a student is negatively correlated with his/her performance. On the other hand, degree of high certainty expressions is positively correlated with the performance. That is, capturing such emotional expressions can be useful for predicting student performance. If such assessment can be done in an early phase of the semester, we may be able to identify who needs what kind of help early on instead of waiting until later in the semester through traditional tests.

In the following sections, we introduce emotional roles in student Q&A discussions. We then present message role classifiers that make use of natural language processing and machine learning techniques. The classified dialogue role information is used in analyzing discussion patterns and related to student performance. Finally, we summarize the current results and discuss future work.

Difficulty Expression in Discussions

It is extremely difficult to devise a category of affect labels given the gradations and subtlety of the way emotions are expressed in language. It is not surprising then that there is no general agreement on how to label affective content and that instead there exist a number of different labeling schemes for different domains (Ordelman &Heylen, 2005). Student messages present additional challenges due to their informal nature and large variances in expressions. However, previous work suggests that at least some affective content can be identified and selected for, independent of context (Kim et al., 2010). For example, in Q&A discussion, certainty categorization was shown to assist in distinguishing between editorial and news writing (Rubin et al., 2006), and may be used to distinguish questions and answers by the presence and absence of confidence.

Identifying a set of categories was an iterative process, and there were three criteria for selection: a) category examples had to be well represented in the corpus, b) researchers had to agree on the categories, and c) categories had to be relevant to student learning. Selection was motivated by the desire to identify students' difficulties and attitudes. We examined discourse that indicated confidence, interest and mastery, urgency, etc. Our final categories were high and low certainty (confidence), tension among participants, and frustration.

Our work takes place in the context of an undergraduate course discussion board that is an integral component of an Operating Systems course in the Computer Science department at University of Southern California. The course is offered every semester, and always taught by the same instructor. The students use discussions, most commonly, to seek help on group-based project assignments. The data collected from eight semesters of this course include 5,056 messages and 1,532 threads from 370 users (180 groups).

Table 1 summarizes emotional expression/role categories that are common and indicate difficulties that arise in the Q&A discussions. Informational roles in Table 2 are directly relevant to question answering behavior: Sink represents messages that request information from others and Source messages provide information for others. Note that answers (Sources) can be given in the form of questions (e.g. "have you looked at the manual?") so Source and Sink are not fully equivalent to question and answer. Emotional roles provide additional information on difficulties while the discussion is being developed. Frustrations often arise due to repetitive interaction without any progress. Tensions capture expressions that show negative emotions or dissatisfaction toward dialogue partners. Certainty categories show the degree of confidence in the posted message.

Category	Description	Examples
Frustra-tion	Expression of hopelessness, anxiety, difficulty, repetitious action without progress	I can't figure it out; :cry: any idea??? give me fault again; I am still confused
Tension	Expression of negative emotional content toward a student/an instructor or by a student /an instructor	You have asked the wrong question; I mentioned (stated, discussed) this in class; The result of this sucks.
High Certainty	Concreteness of Question/Answer with high confidence	You will (must); All you need to do is; Best way; This will be
Low Certainty	Vagueness of Question/Answer with low confidence	I am still confused; Not sure if I understand; I cannot figure out

Table 1: Emotional Roles Q&A Discussions.

Category	Description	Examples
Sink	Requesting information from others	We are getting error. What should I do?
Source	Providing information for others	Have you ever looked at the manual?

Table 2: Information Roles Q&A Discussions.

Data Annotation

Annotating affect involved identifying those speech fragments that reliably indicated an identified emotional role in a repeatable fashion throughout the corpus of student discussion board posts. This was complicated by the highly irregular nature of the message content, which was characterized by frequent misspellings and grammar and syntactical errors, stemming from common parlance, simple carelessness, and Computer Science student language use. This necessitated a high level of selectivity and repeatability in all annotations, as well as reliance on specific patterns of distinct phrases and grammar from within the corpus rather than whole statements.

Among the collection data, we selected two semesters' data, 418 threads that contained 1,841 messages, for the annotation. Human annotators manually marked the emotional and informational roles in the messages until they reached enough agreement on unseen dataset. The annotation iterations over three years involved revising the annotation manual and re-training the annotators.

For inter-annotator agreement, Cohen's Kappa values (Cohen, 1960) were used to measure the final agreement between two annotators. 1 implies perfect agreement while 0 means no agreement between the annotators. 0.7 indicates a good agreement. We compared two annotators' data on 322 messages in randomly selected 30 discussion threads. The Kappa scores for Frustration, Tension, High Certainty, Low Certainty, Sink and Source are 0.92, 0.74, 0.92, 0.95, 0.95, and 0.98, respectively.

Data Processing and Automatic Message Role Classification

Identifying emotional and informational roles of messages are challenging. First of all, there are many ambiguities in expressing feelings in natural language. For example, in the sentence "Specifically, certain threads signaling various things before other threads can wait on these signals thank to the joy of context switching", the poster uses sarcasm in conveying his/her problem in handling context switching. Second, it is also hard to identify true information seeking and providing roles using only surface level features such as a question mark or interrogative words. For example, an answer can be given in a form of a question and vice versa: "Have you checked the manual?" Finally, discussions among undergraduate students are highly unstructured and noisy. Cleaning and preprocessing raw data, transforming them into more coherent data sets, and selecting useful features have been challenging.

The data processing fix common typos and abbreviations, convert contracted forms to their full forms, and transform informal words to formal words. For example, "wats" should be converted to "what is" and "yea" or "yup" are all substituted by "yes." Also, emoticons are replaced by their original meaning (e.g. ":)" was replaced to Emo_SMILE and "–(" was replaced to Emo_CRY). To reduce the variance, we combine personal pronouns (e.g. "I" and "we" were replaced by IWE) and keep their base form (e.g. "is", "was", "are", "were", and "been" returned BE). We also combine "must" and "ought to" into "have to." There are two types of quotes (repetition of previous message content in the post) and programming contents that are dominant in project-based discussions. All the quotes were removed by "google-diff-match-patch" and programming is replaced with CODE tags using regular expressions.

For generating features, we made use of annotators' inputs on the kinds of information that they often use in classifying emotional or information roles of the given message. We generated two types of features: message-level features and thread-level features. The first type of features captures the standard N-grams and emoticons extracted from the message content. Many existing speech act classifiers rely on message-level features including content features in the previous message (Carvalho and Cohen, 2005; Samuel 2000) and we use them for a baseline. The second type of features includes author change, message relative/absolute position, and user role (student or instructor). Also, we include the previous message-level and thread-level features because they can provide context for the current message role. For example, Sources tend to follow Sinks.

Classifiers	Message-level features			Thread-level features
	Unigram	Bigram	Trigram	
Frustration	PROBLEM$_{cur}$ TRY $_{cur}$ STUCK $_{cur}$	SAME PROBLEM$_{cur}$ IWE HAVE $_{cur}$ DO NOT $_{cur}$	FOR SOME REASON $_{cur}$ IWE BE STILL $_{cur}$ STILL DO NOT $_{cur}$	User.Pos $_{pre}$ User.Role $_{pre}$
Tension	GET $_{cur}$ CALL $_{cur}$ DISCUSS $_{cur}$	WHY DO $_{cur}$ NEVER SAY $_{cur}$ DO NOT $_{cur}$	WANT TO DO $_{cur}$ YOU REALLY WANT $_{cur}$ STATE IN CLASS $_{cur}$	User.Pos $_{pre}$ Msg.Pos $_{pre}$ User.Pos $_{cur}$
High Certainty	ONLY $_{cur}$ SHOULD $_{cur}$ SURE $_{cur}$	HAVE TO $_{cur}$ IF YOU $_{cur}$ YOU WILL $_{cur}$	BE PRETTY SURE $_{cur}$ DO NOT HAVE $_{cur}$ NOT WANT TO $_{cur}$	User.Pos $_{cur}$ Msg.Pos $_{cur}$ User.Role $_{cur}$
Low Certainty	IWE $_{cur}$ GUESS $_{cur}$ MIGHT $_{cur}$	NOT KNOW $_{cur}$ DO NOT $_{cur}$ IWE GUESS $_{cur}$	DO NOT KNOW $_{cur}$ IWE BE GUESS $_{cur}$ NOT KNOW IF $_{cur}$	User.Pos $_{pre}$ User.Role $_{cur}$ Msg.Pos $_{pre}$
Sink	?$_{cur,}$? $_{pre}$ IWE $_{cur}$ IWE $_{pre}$	DO IWE $_{cur}$ IWE BE $_{cur}$ CAN IWE $_{pre}$	DO IWE NEED $_{cur}$ WH DO IWE $_{cur}$ DO IWE NEED $_{pre}$	User.Pos $_{cur}$ Msg.Pos $_{cur}$ User.Pos $_{pre}$
Source	? $_{pre}$ IWE $_{pre}$ IWE $_{cur}$	CAN IWE $_{pre}$ DO IWE $_{pre}$ YOU CAN $_{cur}$	DO IWE NEED $_{cur}$ WH DO IWE $_{cur}$ YOU CAN NOT $_{pre}$	User.Pos $_{cur}$ Msg.Pos $_{cur}$ User.Pos $_{pre}$

Table 3: Top Message-level and Thread-level Features based on Information Gain.

Role category	With message and thread features			With message feature only	Role category	With message and thread features			With message feature only
Frustration	Precision	Recall	F1-score	F1-score	High Certainty	Precision	Recall	F1-score	F1-score
J48	0.87	0.87	0.87	0.85	J48	0.81	0.80	0.80	0.78
Naïve	0.76	0.75	0.74	0.74	Naïve	0.73	0.71	0.70	0.70
SVM	0.91	0.91	**0.91**	0.90	SVM	0.84	0.84	**0.83**	0.82
Tension	Precision	Recall	F1-score	F1-score	Low Certainty	Precision	Recall	F1-score	F1-score
J48	0.94	0.93	0.93	0.91	J48	0.88	0.88	0.88	0.86
Naïve	0.87	0.82	0.82	0.79	Naïve	0.73	0.72	0.72	0.70
SVM	0.96	0.95	**0.95**	0.93	SVM	0.91	0.91	**0.91**	0.90

Table 4: Classification Accuracies for Emotional Roles.

Sink	Precision	Recall	F1-score	Source	Precision	Recall	F1-score
J48	0.85	0.84	0.84	J48	0.83	0.83	0.83
Naïve	0.87	0.87	0.87	Naïve	0.87	0.86	0.86
SVM	0.90	0.91	**0.91**	SVM	0.86	0.91	**0.89**

Table 5: Classification Accuracies for Information Roles.

Location / Emotion	Frequency		Average			Frequency		Average		
	Top (%)	Bottom (%)	Relative Location	Absolute Location	Thread Length	Top (%)	Bottom (%)	Relative Location	Absolute Location	Thread Length
Frustration	108(62.4)	65(37.6)	0.52	2.94	5.76	28(53.8)	24(46.2)	0.55	4.42	7.53
Tension	24(23.8)	77(76.2)	0.77	3.69	5.72	6(20.0)	24(80.0)	0.73	3.91	7.90
High Cert.	89(42.2)	122(57.8)	0.63	4.52	7.53	19(40.4)	28(59.6)	0.61	5.83	10.47
Low Cert.	72(51.1)	69(48.9)	0.58	4.16	7.30	14(43.8)	18(56.2)	0.63	4.78	8.13
Sink	213(75.8)	68(24.2)	0.47	2.26	4.91	39(59.3)	30(40.7)	0.51	3.51	6.87
Source	108(26.3)	302(73.7)	0.75	4.04	6.13	32(34.0)	62(66.0)	0.67	5.12	8.60

Table 6: Locations of Difficulty Expressions in Resolved/Unresolved Threads.

Given all the combination of the features, we used Information Gain score for eliminating irrelevant or redundant information. We selected top 2,000 out of 19,465 features generated from the training corpus. Some of the top N-gram features are shown in Table 3. We use the following notation $\text{feature}_{\text{position}}$ where the position represents either the previous or the current message. Most message-level features come from the current message. Some of thread-level features were ranked high, including message/user positions in the thread.

A message can contain more than one role so we chose binary classification. For building the classifiers, we randomly divided 418 threads (1,841 posts) into two datasets: 318 threads (1,404 posts) for training and 100 threads (437 posts) for testing. We then used J48, Naïve Byes, and SVM with RBF in the WEKA package. The training phase was carried out based on 10-fold cross validation. Note that distribution of the positive and negative examples for emotional roles are uneven: e.g. Frustration (261) vs. no Frustration (1,143), Tension (277) vs. no Tension (1,127) in the train dataset. Thus, we performed resampling of the training data toward more balanced distribution: duplicating positive examples and random sampling negative ones. Table 4 presents the accuracies that are compared with the human annotated data in the test dataset. The best F1-scores are highlighted in boldface, ranging from 0.83 to 0.95. The thread-level features indeed seem to help the classification; the F1-scores with both message-level and thread-level features are slightly higher than the ones with message-level features only. The following analyses rely on the classification results.

Difficulty Expressions and Discussion Development

One of the key problems in student discussion assessment is to understand the nature of resolved vs. unresolved discussions and developing strategies for assisting unresolved ones. Also, educators are interested in the characteristics of discussions where students are more engaged and posts that encourage more participation. Here we explore how emotional expressions can be related to discussion thread development or discussion resolution.

Difficulty expressions and discussion resolution

We define a resolved discussion as a discussion in which all of the information seeker's questions get resolved, including initial questions, related questions, similar questions, and questions about derived problems. For the analysis, we used manually annotated data: 189 resolved threads (679 messages) and 32 unresolved threads (159 messages) from two semesters. We examined differences in difficulty expressions using measures of message's location (dichotomous, relative, and absolute message's location in a thread). The results are in Table 6.

We found that more Frustrations were in the top portion in either resolved or unresolved threads, but the percentage of Frustration in the bottom part of unresolved threads is about 9% more than the one in resolved threads. That is, in unresolved discussions, more Frustrations were expressed toward the end of the discussion. Tensions seem to play a similar role in both types of threads, but they appear more in the bottom portion of the threads. High Certainty usually appears later than Low Certainty (in questions) in resolved threads, indicating confident answers. However, unresolved threads end with messages with Low Certainty more often than resolved ones. Obviously, most Sinks are located at the top of threads while more Sources are at the bottom. Overall, the average length of unresolved threads is longer than the one for resolved threads. We conjecture that difficulty of the topic may have contributed.

Difficulty expressions and thread development

Longer threads suggest a high level of engagement among users. To evaluate if emotional and informational roles are related to thread lengths, we built a predictive model using regression techniques. We filtered out very short threads that include 1 or 2 messages, resulting in 733 threads containing 3,805 messages, with 5.19 messages per thread on average. As the dependent variable (thread length) shows a power-law distribution, we conducted a Poisson regression assuming that the logarithm of the expected thread length is the linear combination of emotional and informational role frequencies. As in Table 7, besides Sinks/Sources, Frustrations positively correlated with thread length. Frustration-filled messages may invite more participation. Frequency of High Certainty is negatively correlated, indicating that when there is a confident answer, discussions can get cut short. One of the examples is further participation of students or additional discussions becomes limited when the instructor provides a confident answer. The effect of Tensions seems not significant.

	Dependent Variable (Thread Length)	
Independent Variables	Thread Length	*exp*(Thread Length)
Frustration	$.107^{***}$	1.108
Tension	-0.110	0.982
High Certainty	-1.149^{***}	0.552
Low Certainty	0.024	1.024
Sink	0.547^{**}	1.728
Source	1.758^{***}	5.801

Note: N=733 (threads); $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 7: Thread Length vs. Difficulty Expressions: Poisson Regression Analysis.

Predicting Discussant Performance using Difficulty Expressions

In relating difficulty expressions with participant performance, we made use of correlation and regression analyses. We use project grades as the performance measure since most discussions that we are analyzing focus on class projects. The dependent variable is the normalized project grade, and independent variables are frequencies of emotional and informational roles that the discussant's messages play, as shown in Table 8.

Categories	Project Grade
Frustration	-0.32^{**}
Tension	0.28^{**}
High Certainty	0.21^{*}
Low Certainty	0.14
Sink	0.13
Source	0.29^{**}

Note: N=180 (groups); $^{*}p < .05$; $^{**}p < .01$

Table 8: The Result of Correlation Analysis

The result revealed that four independent variables (Frustration, Tension, High Certainty, and Source) are significantly correlated with the dependent variable. We predict that low performers ask more questions due to their confusion or misunderstanding. However, students who tend to answer others' questions may have understood the topics better, and achieve better grades. As expected, Frustration is negatively correlated to the project grade with significance. Intuitively, students who express more difficulties, without proper assistance, could not have reached good grades. Surprisingly, Tension is positively correlated with higher grades. The students who challenge other discussants could have high levels of engagement in general. We conducted multiple regression analysis to test if emotional/informational roles significantly predicted the normalized project grade. The analysis of variance test suggests that the regression model is significant, $F(3, 176) = 20.75$, $p<0.001$, with 32% variance in student performance being explained by three predictors. The result of the multiple regression is summarized in Table 9.

Variables	B	Std. Error	Beta
Frustration	-0.604^{**}	0.197	-.251
Source	$.874^{**}$	0.254	.168
Tension	$.740^{*}$	0.322	.145

Note: R^2=.32; N=180 (groups); $^{*}p <.05$; $^{**}p<.01$;

Table 9: The Result of Multiple Regression Analysis

Frustration had the largest coefficient with a negative value. This suggests that the more students feel frustrated, the lower grade they get. Such information, i.e. students with high frustration frequencies, can be reported to the instructor for further assistance. Sources still survived while High Certainty was dropped off. Conceptually, Certainty expressions may overlap with Sources to some degree. Lastly, Tension had the smallest coefficient but positively correlated. As described above, we plan to perform further analysis with Tension messages.

Related Work

Existing work on online discussions can provide useful information for instructional analyses of student discussions. Speech act analyses reveal roles that individual messages or participants play (e.g., Jeong et al., 2009). Analyses on argumentation styles (Cabrio and Villata 2012), discussion summarization (Chan et al., 2012; Zhou and Hovy, 2006), and topic mining (Diao et al, 2012) can present an overview of how discussions go. Evaluation of user expertise or contribution quality is also useful for evaluating participant contributions (Chen et al., 2011).

Carvalho and Cohen (2005) present a dependency-network based collective classification method to classify email speech acts. However, estimated speech act labeling between messages is not sufficient for assessing contributor roles or identifying help needed by the participants. We included other features like participant profiles. Also our corpus consists of less informal student discussions rather than messages among project participants, which tend to be more technically coherent.

Requests and commitments of email exchange are analyzed in (Lampert et al., 2008). As in their analysis, we have a higher kappa value for questions than answers, and some sources of ambiguity in human annotations such as different forms of answers also appear in our data. However, student discussions tend to focus on problem solving rather than task request and commitment as in project management applications, and their data show different types of ambiguity due to the different nature of participant interests.

There has also been work on non-traditional, qualitative assessment of instructional discourse (Boyer et al., 2008; Graesser et al., 2005; McLaren et al., 2007), and results have been used to find features for critical thinking and level of understanding. Similar approaches for classifying speech acts in Q&A discussions were investigated in Ravi and Kim (2007). This work captures features that are relevant to analyzing noisy student discussion threads and supports a fully automatic analysis of student discussions instead of manual generation of thread analysis rules.

Finally, there have been studies of student affective states in tutor-tutee dialogue, including boredom, confusion, surprise and frustration. These were analyzed and captured using dialogue states with linguistic features such as cohesion measures (D'Mello et al., 2009). Our work focuses on 'threaded' discussions, and is potentially useful for analyzing student collaborative problem solving.

Conclusion

We have presented a new application of online discussion analysis: supporting pedagogical assessment of online discussions by capturing difficulties expressed. We have described a set of common emotional and informational dialogue roles that individual messages play, and developed a promising classification approach. We have shown that emotional/informational dialogue roles are important factors in explaining discussion development and student performance.

In particular, Frustrations occur more frequently when discussions get longer. The students who express Frustration tend to get lower grades. Such information can be useful for identifying students who need more assistance and alerting the instructor. We plan to combine these results with discussion topics so that we can identify specific areas that need improvement for individual students. Such information can be useful for intervening weak students early on.

Alternative approaches for the classification of emotional or information roles will be explored including unsupervised or semi-supervised approaches that can make use of more data (Jeong et al., 2009). Graphical models can be used for capturing the structure of the conversation (Joty et al., 2011).

There are many research directions that we can pursue. In combination with existing quantitative and qualitative metrics including rhetorical speech acts, coherency, and other linguistic characteristics, the new measures can provide a powerful tool for researchers and instructors.

Acknowledgement

This work is supported by the National Science Foundation, REESE program (award #1008747).

References

Boyer, K., Phillips, R., Wallis M., Vouk M., Lester, J., Learner Characteristics and Feedback in Tutorial Dialogue (2008), ACL workshop on Innovative Use of NLP for Building Educational Applications.

Brown, P. and Levinson, S.C. 1987. Politeness: Some universals in language usage. Cambridge: Cambridge University Press.

Cabrio, E. and Villata, S. (2012). Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions, Proceedings of ACL 2012.

Carvalho, V.R. and Cohen, W.W., (2005). On the collective classification of email speech acts, Proceedings of the SIGIR.

Chen, B.-C., Guo, J., Tseng, B. and Yang, J. (2010) User reputation in a comment rating environment. Proc. KDD 2011.

Cohen, J., 1960. A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, 37-46.

Cohen, W.W., Carvalho, V.R. and Mitchell T.M. (2004). Learning to Classify Email into Speech Acts. Proc. EMNLP.

D'Mello, S., Dowell, N., and Graesser, A. (2009). Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States. Proceedings of the AI in Education Conference.

Graesser, A. C., Olney, A., Ventura, M., Jackson, G. T., (2005). AutoTutor's Coverage of Expectations during Tutorial Dialogue, Proceedings of the FLAIRS Conference.

Graesser, A., VanLehn, K., Rosé, C., Jordan, P., Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. AI Magazine, 22(4).

Jeong, M., Lin, C.Y. Lee, G.G. (2009). Semi-supervised Speech Act Recognition in Emails and Forums, Proceedings EMNLP.

Jeong, A. (2009). The Effects of Intellectual Openness and Gender on Critical Thinking Processes in Computer-supported Collaborative Argumentation. Journal of Distance Education, 22(1), 1-18. 2009.

Joty, S., Carenini, G., and Lin, C.Y. (2011). Unsupervised Approaches for Dialog Act Modeling of Asynchronous Conversations. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011).

Kim, J., Shaw, E., Wyner, S., Kim, T., and Li. J. (2010). Discerning Affect in Student Discussions. In Annual Meeting of the Cognitive Science Society.

Lampert, A., Dale, R., and Paris, C. (2008). The Nature of Requests and Commitments in Email Messages, AAAI workshop on Enhanced Messaging.

Mann, W.C. and Thompson, S.A., (1988). Rhetorical structure theory: towards a functional theory of text organization. Text: An Interdisciplinary Journal for the Study of Text, 8 (3).

McLaren, B. et al., Using Machine Learning Techniques to Analyze and Support Mediation of Student E-Discussions, Proceedings of the AI in Education Conference.

Morgan, B., Burkett, C., Bagley, E., & Graesser, A. C. (2011). Typed versus Spoken Conversations in a Multi-party Epistemic Game, Proceedings of AI in Education.

Ordelman, R. and Heylen, D. (2005). Annotation of Emotions in meetings in the AMI project.

Palmer, S., Holt, D., and Bray, S. (2008). Does the Discussion Help? The Impact of a Formally Assessed Online Discussion on Final Student Results. British Journal of Educational Technology, 39(5):847–858.

Ravi, S., Kim, J. (2007). Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers, Proceedings of the AI in Education Conference, pages 357–364.

Rubin, V., Liddy E., and Kando, N. (2006). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In Computing Attitude and Affect in Text: Theory and Applications.

Samuel, K., (2000). An Investigation of Dialogue Act Tagging using Transformation-Based Learning, PhD Thesis, University of Delaware.

Scandamalia, M. and Bereiter. C. (1994). Computer Support for Knowledge-Building Communities. The journal of the learning sciences, 3(3):265–283, 1994

Searle, J., (1969). Speech Acts. Cambridge: Cambridge Univ. Press.

Zhou, L., Hovy E. (2006).On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.