

An Intelligent Discussion-Bot for Guiding Student Interactions in Threaded Discussions

Jihie Kim, Erin Shaw, Grace Chern, and Donghui Feng

University of Southern California/Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292 USA
+1 310 448 8769

Abstract

Although there are high expectations for collaborative discussion and on-line learning, existing systems for on-line discussion and chat facilities are not fully effective in promoting learning. Pedagogical interventions are often necessary to keep discussions focused and productive. This work focuses on the creation, implementation and evaluation of innovative new software tools for undergraduate engineering courses that promote collaborative problem solving and reflection through automatic scaffolding of student discussions. We describe how discussions in similar past courses can be used to recommend interesting topics for consideration. We also present an approach to modeling student discussions by including an analysis of speech acts. We show how we use the speech act analysis in assessing participant roles and identifying discussion threads that may have confusions and unanswered questions.

Introduction

Our research takes place in the context of the Distance Education Network (DEN) at the University of Southern California where on-campus courses are broadcasted to remote students. The work was motivated by the lack of flexibility and poor assistance capability of the commercial discussion board at USC, and by the potential for utilizing natural language processing (NLP) technology. We are using phpBB, a popular open-source bulletin board with good community support, as the alternative discussion board. The enhanced discussion board we have developed has been used by nine engineering courses.

Although collaborative discussion and on-line learning appear highly promising, our research suggests that existing systems for on-line discussion could be much more effective in promoting learning. Specifically, many student contributions consist of simple question and answer exchanges (Feng et al., 2006a). Elaboration of technical concepts (as opposed to technical details) and reflective dialogue between peers is uncommon.

Discussion threads are often very short, many consisting of only one or two messages, and students do not fully exploit the collaborative problem solving environment to discuss relevant technical issues with one another. The tutor's interventions are often necessary to keep discussions focused. As course enrollments increase, with some introductory courses enrolling several hundred students, the heavier on-line interaction can place a considerable burden on instructors and teaching assistants.

In this paper, we describe an intelligent agent, or *discussion-bot*, that has been implemented within our Discussion Board, which automatically provides responses to student queries. In particular, we show how we model discussion threads using "speech acts". Discussion threads are considered a special case of human conversation, and each post is classified according to speech act categories such as *question*, *answer*, *elaboration* and *correction*. By classifying discussion contributions according to speech act categories, we were able to identify roles that the students and the instructor play in discussions. We developed a set of patterns for analyzing student interactions in discussions. Some of the patterns are used in identifying cases where students may have unanswered questions.

In the following section, we first present the discussion-bot framework. We then introduce the speech act analysis and describe how we model discussion threads using the speech acts. Finally, we present a set of thread patterns we have developed and how they can be used in assessing student interactions.

Background

Discussion-Bot Framework

The discussion-bot framework was created for answering student queries in discussion boards automatically (Feng et al., 2006a). Figure 1 shows the components of the system.

For our study, we analyzed student discussions in an undergraduate Operating Systems course at the University of Southern California. Ninety-eight undergraduate students were enrolled. For the operating systems course, we had two resources available for mining suitable answers to student queries: the supplementary course

documents and threaded discussions from past semesters. Course documents include reading assignments, homework and solutions, project descriptions, and instructions. Archived threaded discussions from previous semesters are also included in the corpus. A threaded discussion includes an initial message together with all responses. All responses are sequentially linked to the original message in chronological order. We use TextTiling (Hearst, 1994) to segment every whole document into semantically-related segments (tiles) and subsequently process tile units (versus document units).

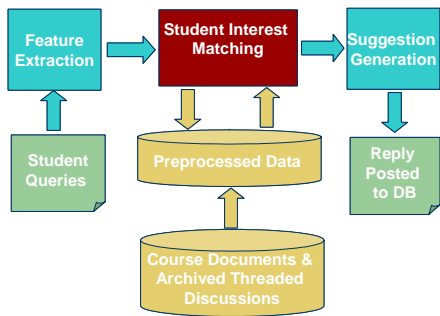


Figure 1: Discussion-Bot components

Student queries often include problem description and answers need explanation. We also found that the discussion corpus from the undergraduate course was incoherent and noisy. Traditional question answering systems in the natural language community usually apply a question-processing module to determine an answer type and extract query terms for the search engine (e.g. Hermjakob et al., 2000; Hovy et al., 2000; Moldovan et al., 2000; Pasca and Harabagiu, 2001). The process is not feasible in our case because it is difficult to discern an exact question and thus identify a single answer type with enumerated query terms. Instead, we retrieve a set of semantically-related passages that match a student's interest by directly computing the cosine similarity between question post and archived data using the TF*IDF technique (Salton, 1989).

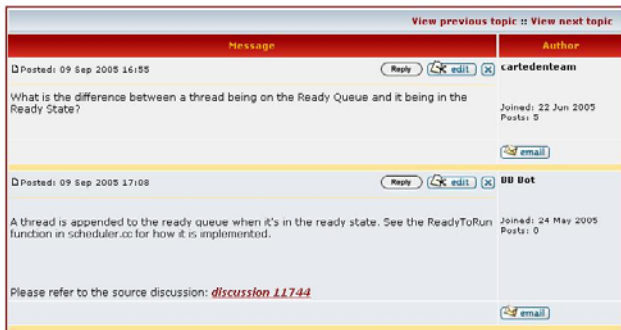


Figure 2. A student query and its Discussion-Bot response.

When a query is posted to the discussion board, the discussion-bot system extracts features from the post first, e.g., the words and word frequencies. Following that, the system tries to match the student's interest in all archived data, both course documents and past discussions. We currently use the cosine similarity between the posted query and archived data using TF*IDF. Intuitively, document tiles and posts with similar words are more likely to be semantically-related. This information is represented by term frequency (TF). However, those with more general terms may be unintentionally biased when only TF is considered, so inverse document frequency (IDF) is introduced to fix the bias. This results in a more general term appearing in more data units with a smaller weight. A list of document segments or message posts related to the question is ranked based on predefined metrics. The suggestion generation module processes the top first candidate in the list based on whether it is a segment of a document or a post. The details on suggestion generation strategies are described in (Feng et al., 2006a). The suggestion is automatically presented on the discussion board, as shown in Figure 2.

Student discussions

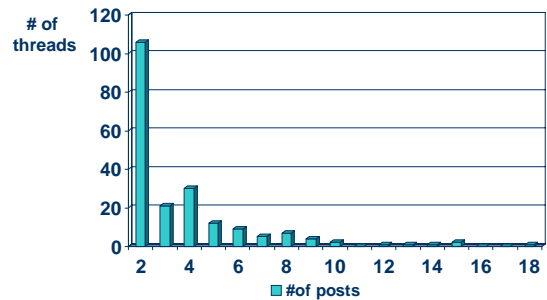


Figure 3: Statistics of Thread Length

Figure 3 shows the distribution of the length of each thread, that is, how many message posts were included in each thread. Although in some cases there are long interactions, up to 18 messages, most threads consisted of only two messages, usually a simple question and answer pair. Elaboration of technical concepts (as opposed to technical details) and reflective dialogue between peers is uncommon. Undergraduate students tend to focus on programming details without necessarily understanding what is to be programmed or how the technical concepts in their lectures are relevant to their assignments. Discussion threads are often very short, and students do not fully exploit the collaborative problem solving environment where they could discuss relevant technical issues with one another. One of the goals of our project is to promote interactions among the students and provide guidance toward more constructive discussions.

Modelling Threaded Discussions with Speech Acts

Unlike in a flat document set, in a threaded discussion each post plays a different role in the discussion. For example, people may make arguments, support or object to points, or give suggestions. However, unlike prototypical collaborative argumentation where a limited number of members take part in the conversation with a strong focus on solving specific problems, online discussions have much looser conversational structure, possibly involving multiple anonymous discussants.

Speech Act Categories

For conversation analysis, we adopted the theory of Speech Acts proposed by (Austin, 1962; Searle, 1969) and defined a set of speech acts (SAs) that relate every pair of messages in the discussion corpus. Though a pair of messages may only be labeled with one speech act, a message can have multiple SAs with other messages. For example a reply message could correct the original message as well as provide an answer. Table 1 gives the specific definitions of each type of post speech act. An automatic classifier is being developed (Ravi and Kim, 2007).

Speech Act	Frequency	%	Cue words
ACK/SUP	56	7.54	"good job" "you got it" "good plan" "good/nice/correct answer" "correct", "thank you"/ "thanks" "i got it" " :)", ";)", "ok"/"okay" "I agree" "its fine with me" "i'm okay with.."
ANNO	3	0.40	"office hours"
ANS/SUG	321	43.2	"perhaps" "how about" "you might", "you probably" "maybe", "try", "i think", "I am/was thinking" "I'm guessing", "my guess" "it should" "it seems" "look at", "check"
CORR/OBJ	41	5.52	"doesn't mean" "are you sure" "what/ how about""didn't work" / "not successful/ "better/ faster/ quicker way"- "i don't think it will work" / "not work" + ... "problem"
ELAB	53	7.13	
QUES	269	36.2	"how" "what" "can we" "are"/ "is" "why" "just/were/was wondering" "I/we have a question" "my question"

Table 1. Speech Act Categories for Individual Messages

We have explored different sets of speech acts based on the assessments of human annotators and found that the above categories are less confusing than other finer or coarser grained categories (Kim et al., 2006). In our corpus, questions and answers comprised the biggest portion of the corpus. This is consistent with the use of the discussion board as a technical question and answer platform for class projects. Figure 4 shows an example of a discussion thread with a sequence of question and answer exchanges. Some

of the QUES and ANS-SUG speech acts are shown in the example.

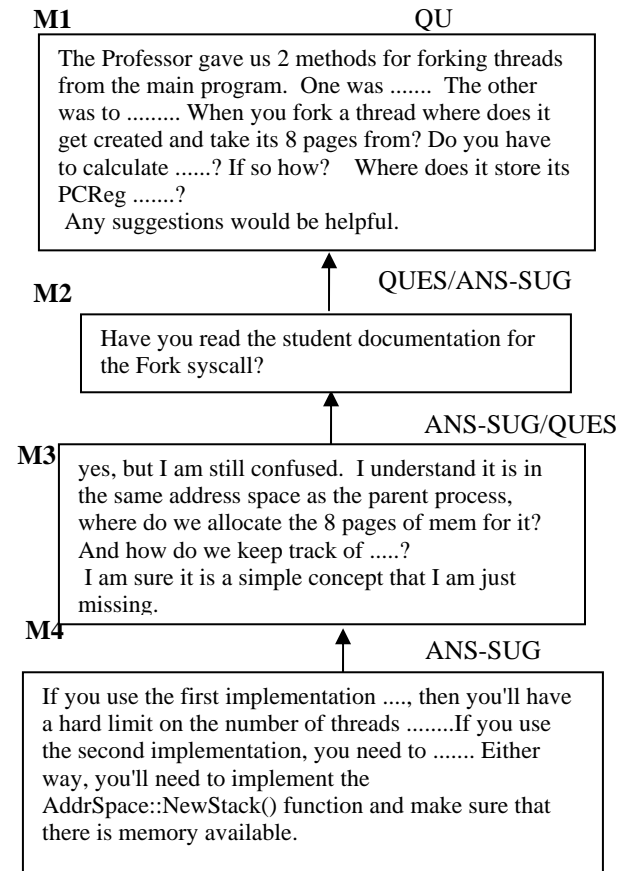


Figure 4: Example of a student discussion thread

Guiding Student Interactions using Speech Act Analysis

This section discusses several approaches for assessing student discussions and guiding the interactions using speech act analysis.

Creating Student Profiles

We investigated differences in speech acts among student discussion contributions. Figure 5 shows the distribution of different speech acts for each group of students. Students were grouped based on the total number of posts they made. As can be predicted from speech act distribution in Table 1, most of the contributions are classified as questions. However, for the students who post many messages, the number of other speech acts, including answers, elaborations, corrections, and acknowledgement increase. These students seem to play more diverse roles in discussions, and their contributions lead to richer

collaborative interactions. Our prior analysis has shown that the students who participate more tend to receive better grades and higher instructor ratings (Kim and Beal 2006).

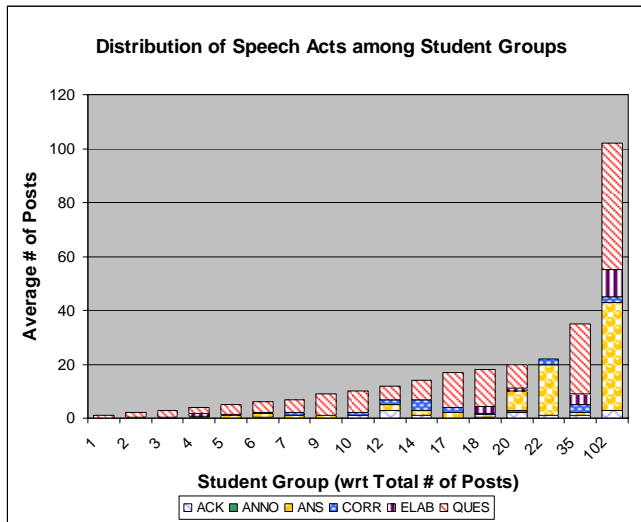


Figure 5: Speech act distribution of different student groups

Effect of Instructor Participation on Student Discussions

The course instructor participated in discussions in many ways; he provided answers directly, gave alternative perspectives, supported student ideas, and elaborated on student answers. Table 2 below shows the frequency distribution of the instructor’s speech acts.

	ACK/ SUP	ANNO	ANS/ SUG	CORR/ OBJ	ELAB	QUES	Total
Instructor SA	18	0	157	5	11	9	200

Table 2: Frequency of instructor speech acts

		Average # of Students	Average # of Student Posts
Threads without Instructor Participation		2.91	3.09
Threads with Instructor Participation		2.60	3.36
Speech Act	ANS + (others)	2.55	3.29
	(others)	3.33	4.50

Table 3: Effects of instructor participation.

Table 3 shows the effect of instructor participation on student participation and thread length. Although threads in which the instructor participated were longer (3.36 vs. 3.09 messages), fewer students contributed to the threads (2.60 vs. 2.91 students). The results are consistent with our earlier findings, including an assessment of a psychology course. However, when we separated the instructor threads in Row 2 by speech act types, specifically, those for which the instructor provided an answer (ANS) versus those for which he provided other scaffolding but no answer, we found that that providing scaffolding without answers had the effect of increasing both the number of students *and* the number of messages in a thread. We are also relating speech acts to discussion quality, to determine, e.g., if instructor participation increases the technical depth of the discussion.

Analyzing Participant Roles and Profiling Threads using Speech Act Analysis

<p>Pattern Group A: short information exchange on non-controversial issues 16 QUES <P1> ANS <P2> 1 ANNO <P1> ACK <P2></p> <p>Pattern Group B: Discussion on somewhat complex issues, answers may have been found. 1 QUES <P1> ANS <P2> CORR <P3> 1 QUES <P1> ELAB <P2> ANS <P2> 1 QUES <P1> QUES <P2> ANS <P1> 1 QUES <P1> ANS <P1> QUES <P2> ANS <P3> 1 QUES <P1> ELAB <P2> ELAB <P3> ACK <P4> QUES <P2> ANS <P4></p> <p>Pattern Group C: collaborative discussion on complex issues, followed by agreeable conclusion 1 QUES <P1> ANS <P2> QUES <P3> ANS <P2> CORR <P3> ACK <P2></p> <p>Pattern Group D: Students may have unresolved issues. 1 QUES <P1> CORR <P2> 1 QUES <P1> ACK <P2> * 1 QUES <P1> ANS <P2> CORR <P1> ANS <P2> ANS <P3> QUES <P2> 3 QUES <P1> ELAB <P1></p>
--

Table 4: Thread profiles: patterns of student interactions without the instructor

To assess student interactions more closely, we analyzed the discussion threads that did not include instructor involvement. We categorized the threads into four pattern groups, A, B, C, and D, as shown in Table 4. Each row in a group represents a thread pattern that belongs to the group. A bracketed identifier is a variable that represents a student. For example <P1> and <P2> means the first and the second contributor, respectively. Multiple instances of an identifier indicate that the student contributed multiple messages to a thread. Patterns in Group A show short interactions for simple information exchange. Group B shows cases where more interactions, such as elaborations, corrections, or questions on answers, were needed to find the answer. Group C includes explicit

agreement among participants. In such cases, students may have a better chance of finding the correct answer. Finally, Group D represents cases where there could be unresolved issues that may need instructor's attention (i.e., there is no answer to the initial question). We found that fully 5 out of the 6 threads in Group D had unresolved issues. In the case marked with *, student <P1> had repeated a question that had been already answered in a previous thread.

Related Work

There have been various approaches to assessing collaborative activities. Various approaches of computer supported collaborative argumentation have been discussed (Shum, 2000). Machine learning techniques have been applied to train software to recognize when participants have trouble sharing knowledge in collaborative interactions (Soller and Lesgold, 2003).

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 detecting human conversation focus. We included other features like participant profiles.

Rhetorical Structure Theory (Mann and Thomson, 1988) based discourse processing has attracted much attention with successful applications in sentence compression and summarization. Most of the current work on discourse processing focuses on sentence-level text organization (Soricut and Marcu, 2003) or the intermediate step (Sporleder and Lapata, 2005). Analyzing and utilizing discourse information at a higher level, e.g., at the paragraph level, still remains a challenge to the natural language community. In our work, we utilize the discourse information at a message level.

Zhou and Hovy (2005) described a method to summarize threaded discussions in a similar fashion to multi-document summarization; but their work does not take into account speech acts. Wan and McKeown (2004) describe a system that creates overview summaries for ongoing decision-making email exchanges by first detecting the issue under discussion and then extracting responses to the issue. Their corpus averages 190 words and 3.25 messages per thread, considerably shorter than the ones in our collection. Marom and Zukerman (2005) generated help-desk responses using clustering techniques, but their corpus is composed of only two-party, two-turn, conversation pairs rather than multi-ply conversation.

There has been prior work on dialogue act analysis and associated surface cue words (Samuel 2000; Hirschberg and Litman 1993). Although they are closely related to our speech act analysis, it is hard to directly map the existing results to our analysis. The interactions in our corpus are driven by problems or questions initiated by students and often very incoherent.

In our previous work (Feng et al., 2006a), we implemented an answer mining algorithm that extract potential answers (the most informative message) using a rule-based traverse algorithm that is not optimal for selecting a best answer; thus, the result may contain redundant or incorrect information. We argue that pragmatic knowledge like speech acts is important in conversation analysis.

Discussion and Future Work

We have presented a new model for assessing pedagogical discourse that is based on an analysis of speech acts within the context of a student discussion board. A machine learning tool based on this model will be used to automatically classify message forms. Our goal is to develop a suite of instructional tools that support the semi-automatic assessment of discussion using diverse measures of contribution quantity and quality. This paper presents several new measures that utilize Speech Act classification: student contribution types, effect of scaffolding, and thread profiling. In combination with existing measures on contribution quantity and technical depth, the measures will help instructors assess qualitative differences among student contributions, how individual students contribute to collaborative problem solving, and how to assess if further help is required.

A fine-grained analysis of discussion activities may help us identify less productive and unfocused discussions where scaffolding is needed and how help should be given. In addition, extensive analysis of student discussion activities and discussion threads can support question answering by extracting useful information from the discussion corpus (Kim et al., 2006).

We are extending the discussion-bot framework, in order to create tools that promote reflective contributions to the discussion by providing two types of recommendations: (a) topics or issues that were considered by students in corresponding graduate-level courses and (b) topics or issues that were considered by students in the same course, previously. In contrast to discussions in undergraduate courses, constructive discussions in comparable graduate-level engineering courses often include insights needed to solve a given problem, analyses of how the problem is similar to or different from ones solved previously, and references to what has been covered in the class. In effect, our system will allow graduate students to become virtual role models for undergraduate students, scaffolding peer interaction and deeper reflection.

Acknowledgement

The work was supported in part by a grant from the Lord Corporation Foundation to the USC Distance Education

Network and in part by National Science Foundation, a CCLI-Phase 2 (Expansion) grant award No. 0618859. We would like to thank Ed Hovy for his suggestions on this work. We also thank Sujith Ravi and Roshan Herbert for their help on the discussion board software.

References

- Austin, J., 1962. *How to do things with words*. Cambridge, Massachusetts: Harvard Univ. Press.
- Carvalho, V.R. and Cohen, W.W. 2005. On the collective classification of email speech acts. In *Proceedings of SIGIR-2005*, pp. 345-352.
- Feng, D., Shaw, E., Kim, J., and Hovy, E.H. 2006a. An intelligent discussion-bot for answering student queries in threaded discussions. In *Proceedings of the Intelligent User Interface Conference*, pp. 171-177, 2006.
- Feng, D., Kim, J., Shaw, E., and Hovy E. 2006b, Towards Modeling Threaded Discussions through Ontology-based Analysis, Proceedings of National Conference on Artificial Intelligence, 2006.
- Hearst, M. A. 1994. Multi-paragraph segmentation of expository text. *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 1994.
- Hermjakob, U., Hovy, E. H., and Lin, C. 2000. Knowledge-based question answering. In *TREC-2000*.
- Hirschberg, J. and Litman, D., 1993 Empirical Studies on the Disambiguation of Cue Phrases, *Computational Linguistics*, 19 (3).
- Hovy, E.H., Gerber, L., Hermjakob, U., Junk, M., and Lin, C. 2000. Question answering in Webclopedia. In *Proceedings of TREC-2000*.
- Kim, J. and Beal, C 2006. Turning quantity into quality: Supporting automatic assessment of on-line discussion contributions, *American Educational Research Association (AERA) Annual Meeting*, 2006.
- Kim, J., Chern, G., Feng, D., Shaw, E., and Hovy, E. 2006. Mining and Assessing Discussions on the Web through Speech Act Analysis, *Proceedings of the ISWC'06 Workshop on Web Content Mining with Human Language Technologies*, 2006.
- Mann, W.C. and Thompson, S.A. 1988. Rhetorical structure theory: towards a functional theory of text organization. *Text*, 8 (3), pp. 243-281.
- Marom, Y. and Zukerman, I. 2005. Corpus-based generation of easy help-desk responses. *Technical Report, Monash University*.
- Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., and Rus, V. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of ACL-2000*.
- Pasca, M. and Harabagiu, S. High Performance Question/ Answering, in *Proceedings of SIGIR-2001*. pp. 366-374.
- Ravi, S. and Kim, J., 2007. Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers, internal project report.
- Salton, G. 1989. *Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.
- Samuel, K., 2000, An Investigation of Dialogue Act Tagging using Transformation-Based Learning, *PhD Thesis*, University of Delaware.
- Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge Univ. Press.
- Shum, B. S. 2000. Workshop report: computer supported collaborative argumentation for learning communities, *SIGWEB Newsl.* 2000., 27-30.
- Soller, A., and Lesgold, A. 2003. Computational Approach to Analyzing Online Knowledge Sharing Interaction, In *Proceedings of AI in Education*, 2003.
- Soricut, R. and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT/NAACL-2003*.
- Sporleder, C. and Lapata, M. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of HLT/EMNLP 2005*.
- Wan, S. and McKeown, K. 2004. Generating overview summaries of ongoing email thread discussions. In *Proceedings of COLING 2004*.
- Zhou, L. and Hovy, E.H. 2005. Digesting virtual "geek" culture: the summarization of technical internet relay chats. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2005.