

Mining and Assessing Discussions on the Web through Speech Act Analysis

Jihie Kim, Grace Chern, Donghui Feng, Erin Shaw, Eduard Hovy

Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292, United States
{jihie, gchern, shaw, donghui, hovy}@isi.edu

Abstract. Online discussion is a popular form of web-based computer-mediated communication, and is a dominant medium for cyber communities in areas of customer support and distributed education. Automatic tools for analyzing online discussions are highly desirable for better information management and assistance. This paper describes an extensive study of “speech acts” in discussions. We present an approach to classifying student discussions according to a set of speech act patterns and show how we use the patterns in assessing participant roles and identifying discussion threads that may have confusions and unanswered questions. We also show how speech act analysis can improve automatic question answering capabilities. This analysis of human conversation via online discussions provides a basis for the development of future information extraction and intelligent assistance techniques for online discussions.

Keywords: on-line discussion board, speech acts, discussion mining.

1 Introduction

With the rapid growth of the Web in the past decade, a large amount of data has been accumulated and may serve as a huge knowledge base for data analysis. Online discussion boards are widely used in areas such as customer support and education, and as a forum for general discussion within cyber communities. People use the forums to collaborate, to exchange information, and to seek answers to problems. For example, in web-enhanced courses, discussion boards are heavily used for question answering and collaborative problem solving (Soller & Lesgold 2003).

Past approaches to mining information from discussion board text, including our own work, mainly focused on finding answers to questions (Feng et al, 2006a; Feng et al, 2006b; Marom & Zukerman, 2005). Most of these techniques simply consider discussion data as text corpus. However, there are increasing needs for modeling discussion activities more explicitly. For example, instructors want to review student discussions in order to understand what kinds of contributions are made by the students and whether they need any assistance or guidance (Painter et al., 2003). In mining answers from discussion corpus, it would be useful to identify which message in a thread contains the most important information (Feng et al., 2006b).

This paper presents extensions to the existing discussion modeling by including an analysis of 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*. In our study, we analyze student on-line discussions in a web-enhanced course. 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 found that students who contributed more to the discussion board played more diverse roles. Instructional scaffolding, i.e., the use of speech acts such as *question*, *elaboration*, and *acknowledgement* had the effect of increasing both the number of student participants and the number of messages posted only when the instructor did not also provide an answer. We also show how we use speech act analysis in identifying discussion threads that may have unanswered questions and need instructor attention. Finally, we present how speech act analysis can improve automatic question answering by detecting the focus of a thread.

2. Approach to classifying discussion contributions according to speech acts

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. The on-line discussion board was divided into six forums, four of which were dedicated to project discussions and two of which were general-purpose administration and knowledge forums. Most of the discussions were about the course projects. The students could start new threads on any of the topics. Most of the discussion threads were initiated by the students.

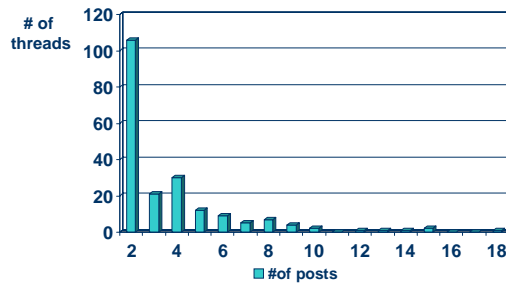


Figure1: Statistics of Thread Length

Figure 1 shows the distribution of the length of each thread, that is, how many message posts were included in each thread. Most threads consisted of only two messages, usually a simple question and answer pair. However, in some cases, there are longer interactions, up to 18 message exchanges.

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.

We found that discussions among undergraduate students are highly unstructured and incoherent. To develop an effective classification approach, we evaluated several different classification methods. In particular, we evaluated agreement between two annotators who labeled each message pair with a given set of speech act categories.

We first used twelve categories listed as Code 1 in Figure 2. We then evaluated two additional coarser-grained sets: 1) directional categories (positive, negative and neutral) and 2) a set that merges similar categories in the original set.

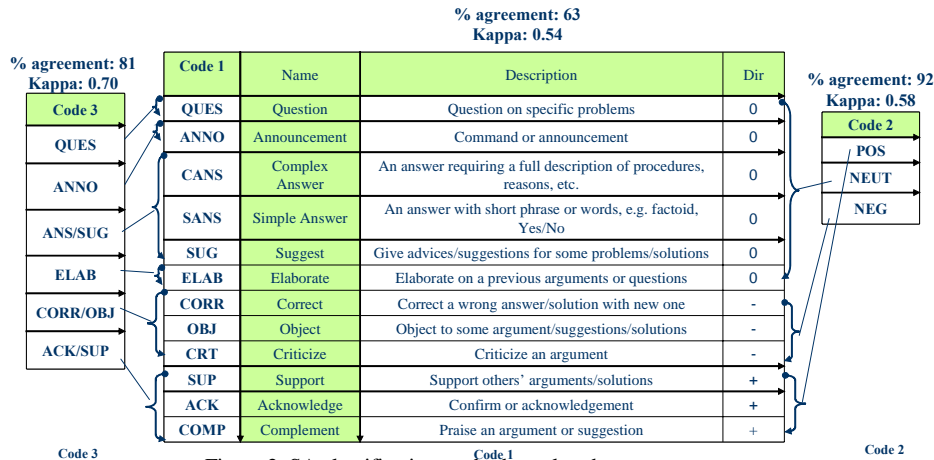


Figure 2: SA classification methods explored.

We used a Kappa measure (Cohen 1960) to verify agreement. Kappa values take into account agreement that can occur by chance.

$$\text{Kappa} = \frac{\text{Observed agreement} - \text{Chance agreement}}{\text{Total observed} - \text{Chance agreement}}$$

As shown in Figure 2, agreement for twelve categories (Code 1) and the directional categories (Code 2) is low. However, by merging similar categories in the original set (Code 3), we were able to obtain a Kappa value of .70 (good agreement) and reduce confusion in category selection. Code 2 fails because of the increase in the chance agreement. We use Code 3 for our analysis below.

Table 1 shows the distribution among the six categories. We found that questions (36.2%) and answers (43.2%) comprised the biggest portion of the corpus. This is consistent with the use of the board as a technical question and answer platform for class projects. The table also lists the surface cue words that are often used during SA classification to discern the speech acts types. We are planning to apply some of the existing machine learning tools for developing an automatic SA classifier with these cue words.

We investigated the relations between two consecutive posts. This can be useful for analyzing how the discussion was conducted among students. Figure 3 summarizes the results. "START" and "END" represent the start and the end of a threaded discussion, respectively. In Figure 3-(a) shows the probability of going from a speech act (SA in left column) to the next speech act (Next_SA in top row). For example, there is a probability of 95.2% that any given discussion will start with a question (QUES). A question often follows by an answer (83.8%). When there is an acknowledgement, the thread often ends with it (73.3%). In Figure 3-(b), each number represents the probability of the previous speech act (Prev_SA in top row) given the current speech act (SA in left column). For example, when the current SA is

CORR/OBJ then there is a high probability that the previous one was ANS (72.7%), although there is low probability of an answer gets corrected (7.4%). These types of relations and cue words seem useful in email classification (Carvalho and Cohen, 2005). We are investigating classifiers that are more appropriate for on-line discussions.

Speech Act	Frequency	Percentage	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	"...and", "also,", "by the way", "same question", "so,"
QUES	269	36.2	"how" "what" "can we" "are"/ "is" "why" "just/were/was wondering" "I/we have a question" "my question"

Table1: Distribution of speech acts and surface cue words

P(Next_SA/SA)	ACK	ANNO	ANS	CORR	ELAB	QUES	END
START	0	0.006	0.042	0	0	<u>0.952</u>	0
ACK/SUP	0.067	0	0	0	0	0.200	<u>0.733</u>
ANNO	<u>1.000</u>	0	0	0	0	0	0
ANS/SUG	0.088	0	0.028	0.074	0.070	0.200	0.540
CORR/OBJ	0.046	0	0.091	0.046	0.046	0.273	0.500
ELAB	0.171	0	0.257	0.029	0.029	0.143	0.371
QUES	0.004	0	<u>0.838</u>	0.018	0.079	0.044	0.018

(a) Probability of next SA

P(Prev_SA/SA)	START	ACK	ANNO	ANS	CORR	ELAB	QUES
ACK/SUP	0	0.067	0.033	<u>0.633</u>	0.033	0.200	0.033
ANNO	<u>1.000</u>	0	0	0	0	0	0
ANS/SUG	0.033	0	0	0.028	0.009	0.042	<u>0.888</u>
CORR/OBJ	0	0	0	<u>0.727</u>	0.046	0.046	0.182
ELAB	0	0	0	0.429	0.029	0.029	0.514
QUES	<u>0.693</u>	0.026	0	0.187	0.026	0.022	0.044
END	0	0.133	0	<u>0.699</u>	0.066	0.078	0.024

(b) Probability of previous SA

Figure 3: Speech act transition probabilities

3. Analyzing participant roles and profiling threads using speech act analysis

We analyzed differences in speech acts among student discussion contributions. Figure 4 shows the distribution of different speech acts for each group of students. Students are 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).

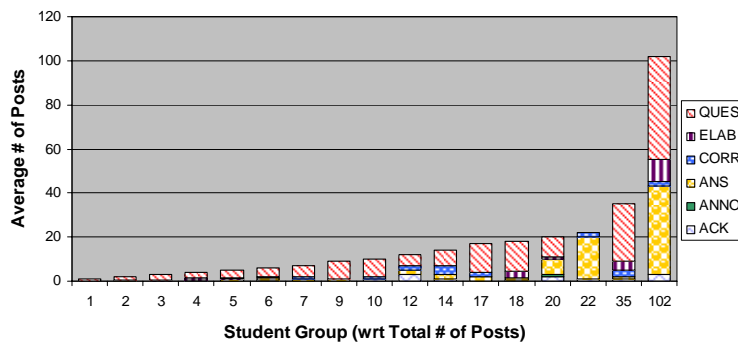


Figure 4: Speech act distribution of different student groups

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 2. 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. When the same student participates more than one time in a thread, the same identifier is repeatedly used. 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 and 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 one of the cases (marked with *), student <P1> was repeating a question that had been already answered in a previous thread.

Pattern Group A: short information exchange on non-controversial issues						
16	QUES	<P1>	ANS	<P2>		
1	ANNO	<P1>	ACK	<P2>		
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	ANS	<P1>	CORR	<P2>		
1	QUES	<P1>	ANS	<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>
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>
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>
3	QUES	<P1>	ELAB	<P1>		

Table 2: Thread profiles: patterns of student interactions without the instructor

4. Improving Questions Answering with Speech Act Analysis: Conversation Focus Detection

In threaded discussions, people participate in a conversation by posting messages. In supporting automated question answering, we want to detect which message in a thread contains the most important information, i.e., the *focus* of the conversation. Unlike traditional IR systems, which return a ranked list of messages from a flat document set, our task must take into account characteristics of threaded discussions.

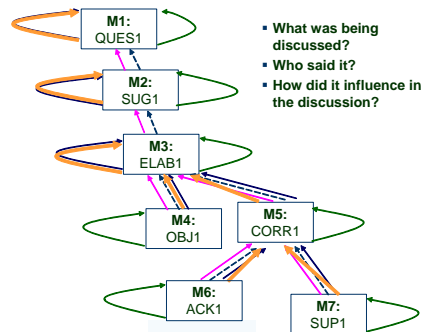


Figure 5. Example Link Generation for Focus Detection.

First, messages play certain roles and are related to each other *by a conversation context*. Second, messages written by different authors may *vary in value*. Finally, since postings occur in parallel, by various people, message threads are not necessarily coherent so the *lexical similarity* among the messages can be analyzed. To detect the focus of conversation, we integrate the above SA analysis, an assessment of message values based on poster trustworthiness and an analysis of lexical similarity. The subsystems that determine these three sources of evidence comprise the features of our feature-based system.

Because each discussion thread is naturally represented by a directed graph, where each message is represented by a node in the graph, we can apply a graph-based algorithm (such as HITS [Kleinberg, 1999] and Page-Rank [Brin and Page, 1998]) to integrate these sources and detect the focus of conversation. The feature-oriented link generation is conducted in two steps. First, our approach examines in turn all the speech act relations in each thread and generates two types of links based on lexical similarity and SA strength scores. We use the cosine similarity between each message pair using the TF*IDF technique (Salton, 1989).

Second, the system iterates over all the message nodes and assigns each node a self-pointing link associated with its poster trustworthiness score. The three features are integrated into the thread graph accordingly by the feature-oriented link generation functions (Feng et al., 2006b).

A speech act may represent a positive, negative or neutral response to a previous message depending on its attitude and recommendation. We classify each speech act as a direction as POSITIVE (+), NEGATIVE (−) or NEUTRAL, referred to as *SA Direction*, as shown in Code 1 of Figure 2. We have used Code 1 for this initial analysis but we are planning to apply Code 2.

▪ Neutral		▪ Positive	
SA	Ws(SA)	SA	Ws(SA)
CANS	0.8134	ACK	0.6844
COMM	0.6534	COMP	0.8081
DESC	0.7166	SUP	0.8057
ELAB	0.7202	▪ Negative	
SANS	0.8281	SA	Ws(SA)
SUG	0.8032	CORR	0.2543
QUES	0.6230	CRT	0.1339
		OBJ	0.2405

Table 3. SA strength scores.

We compute the strength of each speech act in a generative way, based on the author and trustworthiness of the author. The strength of a speech act is a weighted average over all authors.

$$W^S(SA) = \text{sign}(dir) \sum_{person_k} \frac{\text{count}(SA_{person_k})}{\text{count}(SA)} W^P(person_k) \quad (1)$$

where the sign function of *direction* is defined with Equation 2.

$$\text{sign}(dir) = \begin{cases} -1 & \text{if dir is NEGATIVE} \\ 1 & \text{Otherwise} \end{cases} \quad (2)$$

All SA scores are computed using Equation 1 and projected to [0, 1]. For a given speech act, $SA_{ij}(m_i \rightarrow m_j)$, the generation function will generate a weighted link in the thread graph as expressed in Equation 3.

$$g(SA_{ij}) = \begin{cases} \text{arc}_{ii}(W^S) & \text{if } SA_{ij} \text{ is NEUTRAL} \\ \text{arc}_{ij}(W^S) & \text{Otherwise} \end{cases} \quad (3)$$

The SA scores represent the strength of the relationship between the messages. Depending on the direction of the SA, the generated link will either go from message m_i to m_j or from message m_i to m_i (i.e., to itself). If the SA is NEUTRAL, the link will point to itself and the score is a recommendation to itself. Otherwise, the link connects two different messages and represents the recommendation degree of the parent to the child message. Each SA has a different strength score and those in the NEGATIVE category have smaller ones (weaker recommendation).

When we combined poster trustworthiness and SA strength, our best performance on focus detection produced a precision score of 70.38% and an MRR (Mean Reciprocal Rank score) (Voorhees, 2001) of 0.825, compared to human annotations (Feng et al, 2006b). Lexical similarity doesn't seem to be very effective in focus detection. The MRR equation that we use is:

$$MRR_{focus} = \frac{1}{number_of_threads} * \sum \frac{1}{first_correct_position}$$

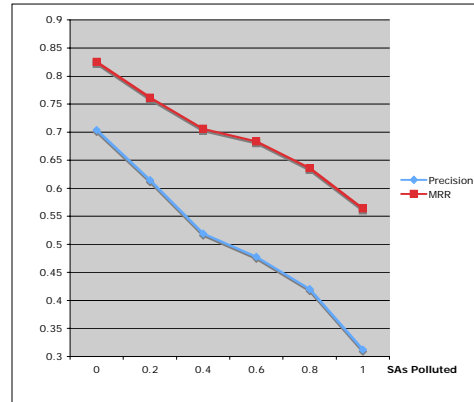


Figure 6. Relation between SA pollution and Precision/MRR

Figure 6 shows the changes in Precision and MRR in focus detection when SA data is gradually polluted by introducing artificial noise.

5. 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 a discussion-bot to automatically answer student queries in a threaded discussion but 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.

6. Summary and Discussion

This paper describes an extensive study of SAs in discussions. By classifying discussion contributions according to SA categories, we were able to identify roles that participants play and develop thread profiles according to SA patterns such as when participants need assistance. We also show how we use speech act analysis in supporting focus detection. We are in the process of extending existing techniques to develop SA classification tools that are appropriate for discussion analysis.

From the perspective of question answering, we present novel techniques to automatically answer complex and contextual discussion queries beyond factoid or definition questions. To fully automate discussion analysis, we must integrate automatic SA labeling together with our conversation focus detection approach. An automatic system will help users navigate threaded archives and researchers analyze human discussion.

The tradeoff and balance between system performance and human cost for different learning algorithms is of great interest. For example we recently proposed a new approach to classifying discussions using a Rocchio-style classifier with no cost on data labeling. In building topic profiles, instead of using a set of labeled data, we employ a coarse domain ontology that is automatically induced from a bible of the domain (Feng et al., 2006c). We are also exploring the application of graph-based algorithms to other structured-objects ranking problems in NLP so as to improve system performance while relieving human costs.

This analysis of human conversation via online discussions provides a basis for the development of future information extraction and intelligent assistance techniques for online discussions, including pedagogical scaffolding capabilities.

References

- Austin, J., 1962. How to do things with words. Cambridge, Massachusetts: Harvard Univ. Press.
- Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107--117.
- Carvalho, V.R. and Cohen, W.W. 2005. On the collective classification of email speech acts. In *Proceedings of SIGIR-2005*, pp. 345-352.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- 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 Intelligent User Interface (IUI-2006)*, pp. 171-177.
- Feng, D., Shaw, E., Kim, J., and Hovy, E.H. 2006b. Learning to Detect Conversation Focus of Threaded Discussions. In *Proceedings of HLT-NAACL 2006*.
- Feng, D., Kim, J., Shaw, E., and Hovy, E.H. 2006c. Towards Modeling Threaded Discussions through Ontology-based Analysis. In *Proceedings of AAAI-2006*.
- Hirschberg, J. and Litman, D., 1993 Empirical Studies on the Disambiguation of Cue Phrases, *Computational Linguistics*, 19 (3).
- 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*.
- Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5).
- Levinson, S. 1983. *Pragmatics*. Cambridge Univ. Press.
- 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*.
- Mihalcea, R. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Companion Volume to ACL-2004*.
- Painter, C., Coffin, C., and Hewings, A. 2003. Impacts of Directed Tutorial Activities in Computer Conferencing: A Case Study. *Distance Education*, Vol. 24, No. 2.
- 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.*, 27-30
- Soller, A., and Lesgold, A., 2003, Computational Approach to Analyzing Online Knowledge Sharing Interaction, *Proc. of AI in Education*.
- 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*.
- Voorhees, E.M. 2001. Overview of the TREC 2001 question answering track. In *TREC 2001*.
- 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 ACL 2005*.