Recognizing Dialogue Content in Student Collaborative Conversation

Toby Dragon[1], Mark Floryan[1], Beverly Woolf[1], Tom Murray[1],

[1] University of Massachusetts Amherst
140 Governors Dr. Amherst, MA USA
{dragon, mfloryan, bev, tmurray}@cs.umass.edu

Abstract. This paper describes efforts to both promote and recognize student dialogue in free-entry text discussion within an inquiry-learning environment. First, we discuss collaborative tools that enable students to work together and how these tools can potentially focus student effort on subject matter. We then show how our tutor uses an Expert Knowledge Base to recognize (with 88% success rate) when students are discussing content relevant to the problem and can correctly link (with 70% success) that content with an actual topic. Subsets of the data indicate that even better results are possible. This research provides solid support for the concept of using a knowledge base to recognize content in free-entry text discussion, and the paper concludes by demonstrating how this content recognition can be used to support students engaged in problem-solving activities.

Keywords: knowledge base, Ill-defined domains, collaboration, inquiry learning

1 Introduction

One of the major challenges facing developers of intelligent tutoring systems for ill-defined domains is maintaining student focus on appropriate content. Students can easily drift from the topic on which they should focus when exploring open-ended environments. A tutor should provide tools to help students maintain proper focus and center their work on the most vital domain content. The tutor should also recognize the content of student work in order to provide appropriate feedback. The introduction of collaboration creates a greater chance that students become sidetracked, but also provides novel opportunities to automatically recognize whether students are engaged in useful learning activities. Here we present research into these concepts of promoting and recognizing discussion of domain content in the collaborative inquiry learning system, Rashi.

Rashi is an inquiry learning system that provides tools and environments necessary for students to have authentic learning experiences by considering real-world problems. The system provides case descriptions for students to investigate, along with information about how to approach each problem [1]. Various data collection methods (interactive images, interview interfaces, video and dynamic maps) provide open-ended spaces for student exploration and acquaint students with methods commonly used by professionals to access and organize information. In the *Human Biology Tutor* (the domain used in our current research), students evaluate patients and generate hypotheses about their medical condition. Patients' complaints form an initial set of data from which students begin the diagnostic process; virtual patients are interviewed about symptoms, Figure 1. Data is made visible by student action (e.g., asking for chest x-rays, performing a physical examination of the patient, or running lab tests). Students move opportunistically from one inquiry phase to another as they sort, filter, and categorize data in order to form hypotheses about the patient's illness.

We have added collaborative tools to this system and demonstrate how we can leverage the information provided by student use of these tools to recognize the domain content on which students are focusing.

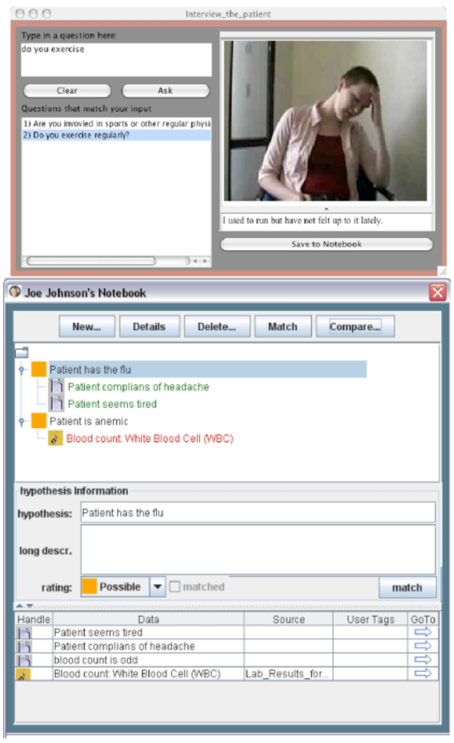

Fig. 1. Students "interview" the patient in the Rashi system (top) and record their hypotheses for a diagnosis along with evidence supporting and refuting that evidence in the notebook (bottom).

There is a large amount of work present in the field of intelligent tutoring systems that is relevant and informative to our current efforts. First, several research groups have explored the potential for collaborative work to improve use of tutoring systems for ill-defined domains. Most prominent among these are the COLER system [2] and COLLECT-UML [3]. Both systems use similar work-sharing techniques to those that we present. These systems give students the ability to view/share work with a team, and offer expert coaching on content. Their coaching does not directly employ their domain knowledge understanding to promote collaborative support; the collaboration and domain-content coaching are separate entities. We hope to take this concept one step further by employing our domain-level knowledge base as a means of improving collaborative efforts. We have also used as reference Soller's research into collaboration that does not attempt domain level support [4]. This work focuses on monitoring the collaborative efforts of students in an attempt to improve on their skills as team members.

Specific to our new line of research in recognizing the content of chat, we must note the large body of research into natural language understanding. Particularly relevant to our work is the research by several groups that have attempted to support students by retrieving related content generated by prior users to provide support. Ravi et al. mine prior student work to find adequate matches [5], while Bernhard and Gurevych analyze the wikiAnswers data store in order to support students [6]. The major challenge to these efforts are data-mining issues; sorting and filtering through large data sets for relevant information. The work presented here differs in that we use a succinct expert knowledge base specifically tailored for the domain and case at hand. This makes the matching of student work tractable and viable without complex NLU techniques, as can be seen by our empirical results. In addition, having ties to our own expert knowledge base can provide new and interesting intervention techniques.

We now describe the collaborative tools added to the Rashi system (Section 2), our system for recognizing content through use of our knowledge base (Section 3), and the empirical studies used to test this system (Sections 4-5). We conclude with a description of future work (Section 6).

2 Collaborative Features

The Rashi system uses collaborative capabilities to enable students to view and share work within a group. These tools allow users to view each others' notebooks, Figure 1, bottom, and to drag and drop both data and hypotheses from others' notebooks to their own. This supports a variety collaborative activities ranging from students working in tightly knit groups, where each student takes on a role and contributes in a specific manner, to students working mostly independently but sharing ideas and thoughts when reaching an impasse. The system also provides a chat facility that enables students to discuss issues with members of their group, Figure 2. Several features, including text coloring, filtering, and new message notifications increase the usability and quality of the discussion tool.

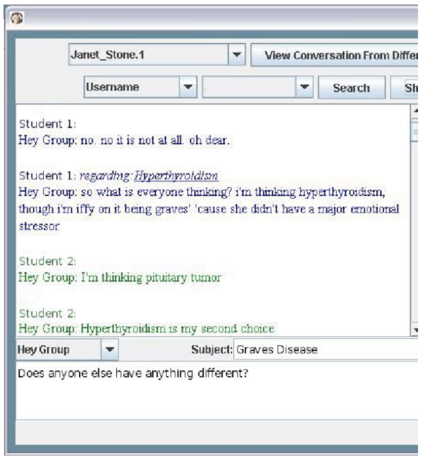

Fig. 2. Collaborative work in Rashi includes student dialogue through chat. Student conversations carry a label (e.g., *Regarding Hyperthyroidism*, middle panel) based on subject (e.g., Graves Disease, bottom panel) typed by a student

Students can create a subject for each message, which allows the team to focus on a specific topic, Figure 2, bottom panel. Chat messages can be filtered by these topics and students can easily respond to the subject by clicking on it. In addition, Rashi allows users a one-click method of automatically setting the subject of a new conversation to the contents of an existing Rashi notebook item. This creates an internal link between the conversation and the notebook item, allowing a confused group member to click on the chat subject and be quickly taken to related work in a group member's notebook.

2.1 Effects of Basic Collaborative Features

We evaluated these collaborative features throughout 2007 and 2008 in both high school and college classrooms and found they caused an increase in the amount of work student completed in the system [7]. Students created more hypotheses, collected more data, and made more connections between their data when collaborative tools were in use. However, we were not able to collect information on whether the increased amount of work was indicative of increased performance. In other words, were the students doing better work, or just more work?

A large study in the Spring of 2009 further investigated whether the amount of work completed within Rashi was indicative of improved performance. We looked for correlations between grade information from teachers and usage statistics from Rashi and found that the amount of work (hypotheses generated, data collected) completed in Rashi was in *no way indicative* of their performance on their final write-ups and overall grade. In other words, seeing an increase in the amount of work students completed was not necessarily an indication of success. We concluded that the major confounding factors involved the issue of students completing a high volume of work, rather than high quality work. Even students who completed a significant amount of work within the system often engage in work that was off-task, repetitive, and/or tangential to the problem at hand, as is often seen when working in ill-defined domains [8]. This result has driven the team's current effort. We built tools to better focus students' collaborative work and use our expert system to automatically recognize the content on which students are currently focusing.

2.2 Content-Focused Collaborative Tools

The chat tool we provided up to this point displays all posts chronologically like other standard chat software. This organization provides an open venue for discussion, but our analysis, as well as others [9], indicates this organization can lead to several forms of noise including the ability to "drown out" other users with a large number of unnecessary intermediate posts. One way to avoid

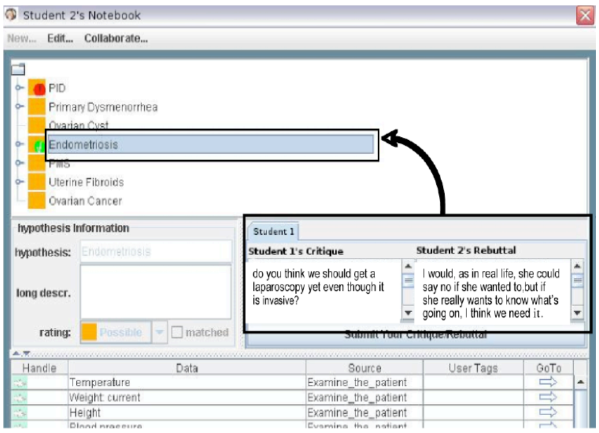

Fig. 3. Student 1 offers a critique of a notebook item (Endometriosis, top panel). The creator of the hypothesis, Student 2, responds to the critique (middle box).

this noise is to provide tools that help students focus on content rather than just providing an open forum. Rashi contains a unique critique-rebuttal feature that fulfills this need and supports students' engagement in topic-oriented discussions, Figure 3. Built into the notebook, this feature enables students to select any item or topic in a group member's notebook and to offer critiques about them. When a critique is given, the owner of the notebook item is notified, and they can respond with a rebuttal, a defense for his or her position, Figure 3, middle panel. This back and forth discussion is by definition focused around subject matter. The two parties can continue to update and resend their critiques and rebuttals, but only the most recent version of each is shown on the interface. This helps to avoid the drowning out of another user. Since this feature is embedded directly into the notebook, it more tightly couples conversations to students' work, thus helping students engage in constructive criticism and organize discussions around content.

3 Recognizing Content

A key feature of the Rashi system is that it provides deep domain understanding by employing the Expert Knowledge Base [1]. Domain experts created knowledge bases for each Rashi tutor using an external authoring tool, thus enabling Rashi to remain domain-independent yet offer content feedback over many different subjects (e.g., biology, forestry, geology). The Expert Knowledge Base is a directed, acyclic graph of domain–related concepts connected with supporting and refuting relationships, Figure 4. At the top of the graph are hypotheses; high-level possibilities of reasonable explanations of the phenomena presented within the given domain. At the bottom of the graph are data or facts and low-level observations about the case at hand. Relationships connect hypotheses and data, sometimes directly and sometimes through mid-level inferences.

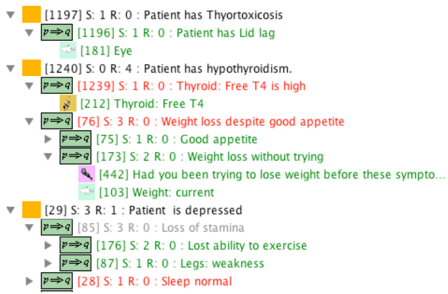

Fig. 4. The Expert Knowledge Base for three diagnoses (thyrotoxicosis, hypothyroidism, and depression) and the evidence supporting (green) and refuting (red) each diagnosis.

Our current work focuses on human biology, presented through differential medical diagnosis. This knowledge base is generalized across all the cases of the domain (e.g., hyperthyroidism, food poisoning, diarrhea) and has been supplemented to suit cases individually, which plays an important role in recognizing content. Using this knowledge base, various types of support can be offered by an automated coaching agent. When the coach recognizes which diagnosis or evidence the student is discussing, it can show supporting or refuting evidence for a student who is stuck, help students create hypotheses that are consistent with their data, or support students to create correct relationships between their hypotheses and data [1]. The addition of collaboration to the system offers the unique opportunity to pair students intelligently by leveraging the knowledge provided by the coaching agent, allowing students to support each other at crucial points in time rather than relying on the coach to handle the entire burden of student support [7].

However, a single coaching agent considering multiple students simultaneously introduces the need for more precise and constantly updated matching of student work to the knowledge base. Automated dialogue content recognition becomes paramount when supporting collaborative efforts, as the tutor needs information about multiple students to provide support for any one. With individualized coaching, the system needed only to reason about a single student at the time support was requested. However, when attempting to intelligently encourage collaboration, the tutor must reason about all students work at the time when any one student needs support. This drives our current research into automated matching of student content and we apply the same matching algorithm to the chat messages between students to as we did to recognize and respond to content for a single student.

4 Empirical Studies of Automated Chat Analysis

Two test groups used the Rashi collaborative system with chat and critique/rebuttal facilities. The first class included approximately 44 middle-school students from a science summer camp at Hampshire College, Summer 2009, and the second group 14 students in an introductory Hampshire human biology course, Fall of 2009. In both studies, students worked with the system for five days and could use the system on their own computers at any time. During these two sessions, a total of 796 non-blank individual chat messages were sent. Each student, however, could only view the chat happening within his or her group of three to five students.

An independent judge (a member of the software team not involved or familiar with the matching scheme or the knowledge base) created a set of comparison data over which the efficacy of an automatic matcher could be assessed. The judge rated all 796 chat messages according to a simple metric: a *content* score of 1 if the message specifically referenced knowledge relating to the case and the student constructively discussed the problem in an effort to perform the desired work. Messages received a score of 0 if they did not contain this material. Through this process, the judge found that twelve percent of the Summer 09 messages and 50 percent of the Fall 09 messages were considered on target in terms of content[1].

The first task was to see how often the recognition algorithm could identify that a given chat message referred to domain content. Thus the automated system gave each message a rating of 1 or 0 for content. We found that overall the system had an 88% success rate identifying messages containing domain content, meaning in 88% of the cases, the automated matching decision and independent judgment agreed, Table 1.

[1] Several students in summer group diluted the chat by saturating conversations with meaningless messages. The summer content score rose to 31% with these messages removed.

Table 1. Data comparison to demonstrate the efficacy of the automated chat-matching algorithm, which reasoned about whether chat content was appropriate for the domain.

Data Set	Total Messages	Judge / Automated Agreement	% Correctly Identified
Summer 2009	496	461	93%
Fall 2009 Case 1	93	84	90%
Fall 2009 Case 2	207	153	73%
Total	**796**	**698**	**88%**

The success rates for identifying the presence of domain content indicates the potential for identifying content within chat messages. However, the system needs to identify precisely what content is being discussed in order to provide the most useful automated help for students. Therefore, rather than converting the match results to a Boolean value, we set the matching algorithm to return the “best fit” content for that message. Once again, the same independent judge compared the student’s statement with this matched content and marked whether the knowledge base entry was appropriate, Table 2.

Table 2. Results of the automated chat-matching algorithm in correctly identifying the specific content of each student message. The judge’s evaluation of the topic automatically identified by the algorithm.

Data Set	Automated Content Matches	Judged Correct Content Matches	% Content Match
Summer 2009	63	44	70%
Fall 2009 Case 1	**69**	**52**	**75%**
Fall 2009 Case 2	45	25	56%
Total	**177**	**121**	**70%**

As seen in the table, in the best case, 75% of the matches found by the algorithm correctly matched student content with exact concepts in the database and the matching success for all messages was 70%.

5 Discussion of Results

A key result of this research is the development of an automatic recognition algorithm that recognizes domain content in student dialogue, a significant milestone in understanding and evaluating student dialogue. This effort, especially the results in Table 2, accomplish the difficult task of matching an infinite space of input (student’s typed responses) with a set of hundreds of knowledge base statements, rather than attempting to predict a binary decision (content or not) as was shown in Table 1. The information provided in Table 2 is significantly more useful for a coaching agent because the tutor can actually understand the precise content that students are discussing. This provides numerous opportunities for coaching, which we discuss in the following section. These results show that a tutoring system can accurately determine the domain content of student dialogue in the majority of cases by using an Expert Knowledge Base. This is a powerful result that implies building knowledge bases is a viable option for the intelligent tutoring community at large when researchers seek to understand the domain content of student discussion.

Another key result of this work is the realization that development of the knowledge base plays a critical role in the success of concept identification. In Rashi, an over-arching knowledge base exists about human anatomy that operates for the domain of differential diagnosis in general. However, the knowledge base is customized on a case-by-case basis (e.g., for hyperthyroidism, food poisoning, diarrhea) in two ways: 1) observational data for each case is different, and so is linked to the database in different ways and 2) the knowledge base is built out for each individual case as needed based on a small generic knowledge structure about the domain (e.g., normal range for blood pressure). This latter feature allows for maximum reusability of the knowledge base, while also minimizing the amount of wasted effort on unused portions of the knowledge base. We put considerable effort into enhancing an Expert Knowledge Base for the cases used in Summer and Fall Case1. In Tables 1- 2, the data used in the Summer and Fall Case 1 were more useful for dialogue recognition than were the data for Fall Case 2. While this leaves us with less-than-ideal results for case 2, it reinforces the idea that our knowledge base structure and creation process are working successfully, since added effort led to direct improvement of matching capabilities.

6 Conclusions and Future Work

Our current work showed success (70-88%) in recognizing content in student discussions. While recognizing content of scenarios is useful, the content is not recognized correctly in the rest of the cases. We also do not currently have a system for identifying when we have potential mismatches, and any such system would not be perfect either. Here we come up against an essential issue of tutoring systems in ill-defined domains. Whenever students are working on authentic, realistic problems and given freedom to work in open-ended environments, coaching systems will have to operate under uncertainty. Therefore, precautions need to be taken to avoid an intervention that would be disruptive or counter-productive if the tutoring system were to be mistaken about the content recognition. We keep this in mind as we go beyond recognition and use this content information to guide students with several different resources available in Rashi. The coach will present both passive information sources relevant to current work, and more active interventions where students are interrupted from potentially wasteful behavior and prompted to re-focus on content by discussing with group members.

An example of our more passive support techniques is a “Suggested Topics” list adjoining the chat window. The system will populate the list with items that are related to a group’s current work according to the Expert Knowledge Base. The group can then see the connections and gaps in their collective work. Clickable links can be automatically generated and will quickly show students what elements within the tutor are relevant to the domain knowledge associated with their current conversation. By connecting the system in this way, students can quickly move through tutor elements that are related to the chat, and thus students will have swifter accesses to resources relevant to their discussions. These suggested topics and related links could be updated dynamically while the system continuously watches the collaborative effort of the students.

Content recognition can also provide active interventions in Rashi. Active interventions require that the system detect a specific moment at which an intervention should be given. What opportunities will be recognized as appropriate times for prompting students to discuss the argument with one another? Some likely candidate situations include:

- One student is missing supporting or refuting arguments that another student has identified
- Two students have partially completed different parts of the same argument structure
- One student has a hypothesis, and another student has data to support the hypothesis, but neither has formed an argument yet.
- One student has a supporting argument and another student has a refuting argument for the same hypothesis

A pilot study of this type of intervention was conducted along with the Fall 2009 experiments. The system did not yet have automated matching capability, so students were asked to match their statements manually to the knowledge base in order to receive feedback. The pilot was attempted one time for the 14 students involved, and attempted to recognize the first situation listed above. The system did find this scenario and alerted the collaborating students in five individual cases. Out of the five cases, one student did use the given data to support their argument. The use of the system was so limited that we cannot make any real generalizations from this pilot, but we did recognize anecdotally the potential for such a system, and the obvious need for an automated matching scheme that would allow for such interventions on a regular basis.

In order to make these interventions as accurate as possible, future work will also include improvements to the content recognition algorithm by including context in the search process. The automatic recognizer often finds a list of possible subjects for a given chat message. In our experiments we simply used the top match. However, context clues can inform an automatic process when the top match may not the best choice. The first and foremost of these context clues is *temporal proximity*. Currently, each message is considered independently when seeking a link with the Expert Knowledge Base. However, from our data we see that it is likely the subject of a given chat message is the same or similar to the messages around it. An automatic recognizer can weight potential matches by considering the matched content around the given message.

Another major context clue that has not been exploited in the current matching algorithm is *case-specific proximity*. We also noted in our data analysis that large portions of the chat refer to knowledge that is unique to the current case. For example, if the patient in a particular Rashi case is allergic to bees but not to anything else and students are chatting about allergies, then it is more likely that they are specifically referencing bee allergies (and not cat allergies, for example). Again, we can use this type of information to weight our matching choices, helping us more accurately pinpoint the specific knowledge base components related to student discussion.

Acknowledgements

This research was funded by an award from the National Science Foundation, NSF 0632769, IIS CSE, Effective Collaborative Role-playing Environments, (PI) Beverly Woolf, with Merle Bruno and Daniel Suthers. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Dragon, T., Woolf, B. P., Marshall, D. and T. Murray. Coaching within a domain independent inquiry environment. Fifth International Conference on Intelligent Tutoring Systems. Jhongli, Taiwan, Springer 4053, 144-153 (2006)
2. Constantino-Gonzalez, M. A., Suthers, D. D., Escamilla de los Santos, J. G. Coaching web-based collaborative learning based on problem solution differences and participation. International Journal of Artificial Intelligence in Education 13(2-4), 263-299. (2003)
3. Baghaei, N., Mitrovic, A. From Modelling Domain Knowledge to Metacognitive Skills: Extending a Constraint-Based Tutoring System to Support Collaboration, 11th International Conference on User Modelling, UM 2007 4511, 217-227 (2007)
4. Soller, A., Martinez-Monez, A., Jermann, P., Muehlenbrock M. From mirroring to guiding: A review of state of the art technology for supporting collaborative learning." International Journal of Artificial Intelligence in Education 15(4): 261-290 (2005)
5. Ravi, S., Kim, J., Shaw, E. Mining On-line Discussions: Assessing, Technical Quality for Student Scaffolding and Classifying Messages for Participation Profiling. Educational Data Mining Workshop for the Conference of Artificial Intelligence in Education. 70-79. Marina del Rey, CA. USA. July (2007)
6. Bernhard, D., Gurevych, I. Answering learners' questions by retrieving question paraphrases from social Q&A sites. Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications for the Association for Computational Linguistics Columbus, Ohio: 44-52 (2008)
7. Dragon, T., Woolf, B. P., and Murray, T: Intelligent Coaching for Collaboration in Ill-Defined Domains. Conference of Artificial Intelligence in Education, Brighton, England: 740-742 (2009)
8. Lu, J., Lajoie, S. P. Facilitating medical decision making with collaborative tools. In Proceedings the World Conference on Education Multimedia, Hypermedia & Telecommunications (pp. 2062-2066). Norfolk, VA: AACE (2005).
9. Truth Mapping: A Tool To Elevate Debate: http://www.truthmapping.com