Exploring the Effectiveness of Social Capabilities and Goal Alignment in Computer Supported Collaborative Learning

Hua Ai, Rohit Kumar, Amrut Nagasunder, Carolyn P. Rosé

Language Technologies Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213
{huaai, rohitk, cprose} @ cs.cmu.edu

Abstract. In this study, we describe a conversational agent designed to support collaborative learning interactions between pairs of students. We describe a study in which we independently manipulate the social capability and goal alignment of the agent in order to investigate the impact on student learning outcomes and student perceptions. Our results show a significant interaction effect between the two independent variables on student learning outcomes. While there are only a few perceived differences in the student satisfaction and the tutor performance as evidenced in the questionnaire data, we observe significant differences in student conversational behavior, which offer tentative explanations for the learning outcomes we will investigate in subsequent work.

Keywords: social interaction, conversational agents, collaborative learning

1 Introduction

Much prior work demonstrates the advantages of group learning over individual learning, both in terms of cognitive benefits as well as social benefits [1][2]. From a cognitive standpoint, one can argue that a major advantage to learning in a group is that when one is exposed to an alternative perspective, it provides the opportunity to question one's own perspective, which in turn offers an opportunity for potential cognitive restructuring. In order to achieve this benefit, a major emphasis of work on scaffolding collaborative learning [3] has focused on drawing out aspects of an issue where there is a disagreement between students so that they will address the disagreements explicitly and benefit from that negotiation process. In line with this, work on formalizing the process of collaboration in order to identify events that are valuable for learning has in many cases focused on formalization of argumentation [4]. In this paper, we again investigate how conflict and negotiation relate to learning, however instead of viewing conversation in terms of individual events or sequences of events that occur and can be counted, we take a more abstract approach and instead characterize spans of text as exhibiting a bias towards one stance or another. In this way, we are able to abstract away from individual actions and think at the level of bias and influence within a discussion, quantifying bias in the same way but adopting different units of analysis in order to view bias and influence at different grain sizes. As a methodological contribution, we discuss how we use as a tool for quantifying bias a state-of-the-art topic modeling technique from the field of language technologies referred to as ccLDA [5].

We begin with a classroom study of collaborative engineering design where students work in pairs on the design of a power plant. This learning task involves negotiating between two competing objectives. Specifically, one student in the pair is assigned to the goal of maximizing the power output of the power plant. The other student, in contrast, is assigned the goal of minimizing the negative environmental impact of the design. Similar to our prior studies of collaborative learning [6][7], a conversational agent participates with the students in the design task in order to provide support. The unique contribution of this study is that we explore the introduction of bias in the way the agent presents information towards one student's stance or the other. In addition to investigating the effect of the manipulation on learning, we investigate the extent to which students are sensitive to displays of bias in the language of their human partner and that of the agent, to what extent the agent's displayed bias affects the bias displayed by the individual students, the interaction between the individual students and the agent, and finally the interaction between the pair of students themselves. As an independent factor, we also manipulate the extent to which the agent exhibits social behaviors designed to build solidarity with the agent in order to investigate whether these solidarity building behaviors either magnify or dampen the effect of the bias manipulation.

In the remainder of the paper we first review the literature on the connection between conflict and learning. We then describe our experimental study. Next we explain our methodology for measuring bias. We then detail our results. We conclude with discussion and directions for future work.

2 Previous Work

Previous work on building socially capable conversational agents focuses on designing social interaction strategies, which fall into the category of social interface in the taxonomy proposed by [8]. The goal of these designs is to enable users to interact in an intuitive and natural way with the agent to perform intended task. For example, Morkes et al [9] implemented a task-oriented conversational agent which uses preprogrammed jokes. They show that this humor-equipped agent is rated as better and easier to socialize with by human participants. In another line of work, Wang and Johnson [10] found that learners who received polite tutorial feedback reported higher increase in self-efficacy at the learning task. Social strategies are also found to be effective in multi-party conversations, such as in computer supported collaborative learning. Higashinaka et. al. [11] found that an agent's use of emphatic expressions improved user satisfaction and user rating of the agent. In general, computer agents which are friendly and helpful to users are favored.

In a multi-party conversation between students and a computer tutor, it is important to build social connections between students and the computer tutor so that students can enjoy the learning process and feel more positive about themselves. At the same time, it is also important to verify that these connections have a positive impact on students' task-related behaviors since such conversation is highly task-oriented. Previous work [12] finds that alignment is a strategy that people use in human-human conversations to complete tasks because they believe it is beneficial in helping both interlocutors to reach mutual understanding. Reeves and Nass [13] further show that although people do not generally believe that computers have human minds, they still behave in similar ways to computers than to fellow humans. Given these two results, we would expect to see people align to conversational agents in human-computer interactions too. This is actually confirmed by a series of studies conducted by Nass and his colleagues [14][15]. They find that human users align to conversational agents at both lexical (semantic) and syntactical levels. The level of the alignments depends on the users' believes on their conversational partners' perceived competence.

Due to the restricted syntax structure presented in our tutoring conversation, in this study we only focus on examining lexical alignments. We explain in Section 4 how a ccLDA model is used to measure the bias of a student's stance in terms of the topics covered in user utterances. These topics are later used to measure the alignment of student utterances at the lexical (semantic) level.

3 Method

We are conducting our research on dynamic support for collaborative design learning in the domain of thermodynamics, using as a foundation the CyclePad articulate simulator [6] which allows students to implement design ideas using graphical interface widgets, and to explore the relationships between the settings of various parameters within the cycle design. In the collaborative design exercise described below, students work in pairs to struggle with trade-offs between power output and environmental friendliness in the design of a Rankine cycle, which is a type of heat engine.

106 Students participated in the study by attending one of six lab sessions, which were structured into multiple phases during which we strictly controlled for time. At the beginning of each lab session, students were lead through formal training on the simulation software from an instructor using power point slides and the Cyclepad simulation environment. They then worked through optimizing some Rankine cycles in Cyclepad using information from a booklet given to them, which was developed by a professor from the Mechanical Engineering Department. Subsequent to this, they took the pre-test, immediately before the experimental manipulation. The exploratory design exercise, which followed, was where the students worked in pairs using CyclePad and the ConcertChat collaboration environment [16]. Students were instructed that they should negotiate with their partner in order to meet their own assigned design objective, namely either to maximize Power output (in the Power condition) or to minimize environmental impact (in the Green condition). This collaborative design exercise was followed by the post-test and the questionnaire and finally a closing activity in which the student was able to work independently with CyclePad to improve the design they developed with their partner. We assigned each student within each pair to a different competing goal, with one student instructed to

increase power output as much as possible and the other student instructed to make the design as environmentally friendly as possible. The trade-offs involved in this task offer students the opportunity to find one of the major motivations for seeking to increase the efficiency of a designed cycle.

During the interaction, students use a collaboration software package called ConcertChat [16] to chat with each other in pairs as well as using the digital whiteboard associated with that environment to pass graphical information back and forth to one another. In all cases, a tutor agent participated with the students in the chat. The experimental manipulation only affected how the tutor agent behaved. In all other respects, the experience of students in all conditions was the same.

The experimental manipulation was a 3X3 between subjects experimental design. For the first independent variable, we contrast 3 social conditions (No Social, Low Social, and High Social) where dialogue agents present different amounts of social behavior within the chat environment. Our dialog agent exhibits three different types of positive social-emotional behavior: showing solidarity, showing tension release and agreeing. In most cases, these strategies are realized by prompts that appear in the chat. The frequency of social behavior in our socially capable tutors is regulated using a parameter that specifies the percentage of tutor turns that can be social prompts. The tutors used in the high and low social conditions differ only in the setting of this parameter. Specifically, in the case of low social tutor, the threshold parameter is set to 15%, i.e., for every 100 turns the tutor says almost 15 that were generated by the social interaction strategies. The high social tutor was configured to generate up to 30% social prompts. In the non social condition, no social behavior is realized.

For the second 3 level independent variable, we design 3 conditions in which the dialogue agents show different levels of support (Yes-Match, No-Match, and Neutral) to an individual student. In the Match condition, the dialogue agent shows a bias towards the student's design goal in how it presents information. In the Mismatch condition, the dialogue agent shows a bias towards the student's partner's design goal. In the Neutral condition, the dialogue agent does not show any bias towards either student's stance. In all cases, the information presented by the tutor is the same. The only difference is the bias exhibited. For example, where the Green biased tutor might say "What is bad about increasing the heat input to the cycle is that it increases the heat rejected to the environment." The neutral tutor would simply say "Increasing hear input to the cycle increases the heat rejected to the environment."

As outcome measures, we examined learning gains between Pre and Post test. 35 multiple choice and short answer questions were used to test analytical and conceptual knowledge of Rankine cycles. We analyzed the conversational behavior in the chat logs. Finally, we compared answers to affective questionnaire items across conditions.

4 Modeling Conversational Dynamics

In this study, we measure the bias of a system/user utterance towards one stance or another by applying a topic discovery model on our tutoring dialogs [5]. LDA models have been widely used to discover topics on large collections of unannotated data [17] using lexical features by modeling the word distributions represented in the data. For example, it has been used to predict responses to political webposts [18], to study the history of different research fields [19], and so on. What is unique about our application of this technology is that we apply it to conversational data for the purpose of modeling how users are interacting with each other. For each utterance, we compute a score to represent to which degree the utterance displays a bias towards one perspective or another.

In our study, we apply a cross-collection Latent Dirichlet Allocation (ccLDA) model [5], which is a variant of the LDA model. With the original, simpler version of LDA, it would be difficult to model how the same topics might be represented differently by speakers representing different points of view. What the ccLDA model has to offer is the ability to build in a level of representation referred to as a collection. Corpora are composed of collections of documents. ccLDA will construct a topic model where each topic will have a separate version for each collection, where those collection specific topic models will represent what is distinct about how those topics are expressed within that collection, as well as a background model, where the same topics again are represented in terms of what is common across collections. A model with this structure is about to be used to compare multiple text collections by capturing similarities and differences across them. Since the two students who participate in each pair are assigned different objectives at the beginning, it is intuitive to apply the ccLDA model to model how the students in the two different conditions discuss similar topics, but express a different point of view through those topics.

TOPIC 1			**TOPIC 2**		
Background	**Green**	**Power**	**Background**	**Green**	**Power**
Heat	11000	yah	power	low	generates
quality	values	blades	decreases	500	makes
right	different	sir	nuclear	12800	85
max	makes	dunno	make	sort	different
decrease	larger	kk	85	1	7000
possible	graphs	x85	cycle	tutors	12000
goes	bit	rejected	work	effeciency	qdot

Table 1: Topics Extracted from ccLDA

To use the ccLDA model, we first separate our dialog data into three collections: those turns that were contributed by the student in the Green condition, those turns that were contributed by students in the Power condition, and those turns contributed by the tutor agent. Our ccLDA model has two collections, namely, a Green collection and a Power collection. We do not include the tutor turns within either collection. When we apply ccLDA to this corpus, then, we get three different topic models, namely, one associated with the Green perspective, one associated with the Power perspective, and one background model representing what is common between the two. When applying ccLDA, one must set a parameter for the number of topics. Because our corpus is relatively small, we set this value to 2. Thus, in all three models, we have the same 2 topics, where a topic is defined as a distribution of words, where the probabilities represent the strength of association between the word and the topic within the model. Table 1 gives an example of the top 7 words selected for each data collection for the two topics.

We designed three metrics for estimating bias towards either the Green perspective (G) or the Power perspective (P) using our ccLDA model. An example where these metrics are applied is presented in Table 2.

Author	**Text**	G_Max	P_ Max	G_Avg	P_Avg	G_Wt	P_ Wt
Stu1	whats ur goal?	0	0	0	0	0	0
Stu2	green as possible	1	0	0.5	0	0.5	0
Stu1	mine is generates the most power	0	2	0	2	0	2
…							
Tutor	If you increase the maximum temperature (T @ S2) of the cycle, what happens to the cycle efficiency?	1	0	0.5	0	0.5	0
Tutor	Cycle Efficiency improves by increasing Tmax.	0	0	0	0	0	0

Table 2: Three types of topic associations

Max Topic-word association (G_Max and P_Max). In the Max Topic-word approach, for each collection specific model we compute a score for each topic, where we count the number of words in the list of the N most strongly associated words with that topic in the corresponding model. The largest number identified for any one topic within that collection specific model is the score for that collection. In this way, we can compute a score for each perspective, since there is one collection specific model per perspective. Hence, for a piece of text that has 2 terms matching with Topic 0 of Green and 1 term with Topic 1 of Green, we would consider 2 to be the score for Green. By averaging over all contributions for the same student within a conversation, it is possible to use this metric to get an average Max Topic-word score for each student for each perspective.

Average Topic-word association (G_Avg and P_Avg). In the Average Topic-word approach, we find the topic-word associations, as in the previous approach. But here, we average scores across topics within a collection specific model rather than choosing the maximum value.

Weighted Topic-word association (G_Wt and P_Wt). This is a heuristic approach that is similar to the previous approach but which uses the weights of topic terms provided in the ccLDA probability model distributions. Whereas in the Average Topic-word approach, each word contributes 1 to the topic specific sum we compute for each topic in each collection specific model, here we add a weight that is computed by multiplying the weight of the term within the background model with the weight of that word in the topic within the collection specific model. We observe that the background models prioritize important, domain-specific terms by giving higher weights. Hence, in this approach, we consider the product of the weight of a word in the Background model and its weight in the specific collection so that the relative weighting of domain important terms is more important for the final weight than terms that are less important for the domain.

We validated the metrics by verifying that students in the Green condition were assigned higher Green bias scores than students in the Power condition, anda students in the Power condition were assigned higher Power scores than students in the Green condition. This was true in all cases, although the differences were only statistically significant for the first two metrics. All three metrics were highly correlated, with R values between .68 and .99. We further validated the metrics using data from a questionnaire where students were asked to rate their partner based on how hard they perceived that their partner attempted to build an environmentally friendly power plant. There was a significant negative correlation between the first and third metrics used to compute a Power score for the partner's conversational contributions and this question's numeric value, and a marginal negative correlation in the case of the second metric.

5 Results

5.1 Learning Outcomes

Recall that our experimental manipulation was composed of two independent factors, which we refer to here as Social (No Social, Low Social, and High Social) and Match (Yes-Match, No-Match, and Neutral). We first look at the most important evaluation standard in tutoring applications – the student learning gains. Using an ANCOVA with Objective Post-test as the dependent variable, Objective Pretest as a covariate, and Social and Match as independent variables, and Session as a random variable, we determined that there was a significant effect of the Social Manipulation (F(2,94) =

5.27, p < .01) where the Low Social condition was significantly better than the other two with an effect size of .83 standard deviations in both cases. There was a marginal interaction between Social and Match F(2,94) = 2.57, p = .08, where Low Social is only significantly better than the other conditions in the case where Match is Yes-Match. All other combinations of Social and Match were statistically indistinguishable.

In general, students learn the most in the condition with the tutor that showed a bias towards their design goal (Yes-Match) and Low Social. Based on the interaction effect between Social and Match, we believe that it is important for the computer tutor to not only establish social connections with the students, but also been viewed as supportive of the students' objectives in order to maximize students' learning outcomes.

5.2 Questionnaire Data

We then look into the questionnaire data to see whether the students perceive the social and goal manipulation we designed in this study. Using an ANOVA for each questionnaire question as dependent variable and Social and Match as independent variables, we determined that there was a marginal effect of Match on rating of tutor as supporting the student's objectives (F(1,102) = 2.77, p = .09), where the tutor was seen as supporting students marginally more in the case where the goals matched. There was no effect of either variable on the perception of whether the tutor supported the partner's goal.

The effect of the Match manipulation was demonstrated in other aspects of the experience, however, according to the questionnaire. For example, on the questions designed to assess the extent to which a student's partner influenced their perspective as a result of the conversation, we observed a significant interaction effect between the Social manipulation and the Match manipulation, such that when the tutor did not exhibit any bias, there was no significant effect of the Social manipulation, but with either agent that showed a bias, either matching the student's bias or the partner's bias, the High social condition significantly reduced the perceived influence of the partner's perspective. Using our bias detection approach, we determined that students were significantly more distinct from their partner in terms of measure of bias in the case where the tutor showed a bias towards one perspective or another, thus magnifying the contrast between the students. This could explain the pattern of behavior we see here. In the case of the neutral tutor, the polarization was less, so the dampening effect of the Social manipulation would not be felt as strongly.

5.3 Conversation Data

Apart from questionnaire data, we can observe an effect of our experimental manipulation on conversational patterns. We have already discussed effects related to bias in the conversation. Here we measure the extent to which students were sensitive to the social aspects of the tutor's behavior that we manipulated through our two independent variables. We began by manually classifying student turns into three categories:

- AboutSocial – student turns on social behaviors, including greetings, farewell, smiling faces, rude words, jokes
- Offtask – student turns talking about off-task topics, like weekend plans, etc
- AboutTutor – student turns that make negative comments about the tutor

We compute the number of AboutSocial, Offtask, AboutTutor turns for each student. Using an ANOVA for each of the three categories as dependent variable and Social and Match as independent variables, we observe that there is a significant effect on AboutSocial (F(2,29)=9.91, $p<0.0001$), where a student's social behavior is significantly lower in the condition with the No Social tutor than with the Low Social tutor (with an effect size of 1.8) and High Social tutor (with an effect size of 2.0). Similarly, there is a significant effect on Offtask (F(2,97) = 3.30, $p < .05$), where students engage in more off task behavior in the No social condition than in the Low and High social conditions (with effect size of .35 standard deviations in both cases). We also observe a significant effect on AboutTutor (F(2,97) = 5.74, $p < .005$), where students utter more negative comments about the tutor in the High social condition than in the Low Social condition (an effect size of 1.1) and the No Social condition (an effect size of 1.28).

Based on our results, we suggest that students will show more social behaviors and focus more on the task when the tutor shows social behaviors. However, when the tutor performs too much social behavior, the students get distracted and start to make fun of the tutor. This is in addition to the dampening effect of the influence students were perceived to have on one another in the High social condition.

6 Conclusions and Current Directions

In this paper we have described an investigation into the issue of competing biases or stances, and how their presence in a conversation, from human or computer participants, affects the learning, interactions, and perceptions of the encounter. Specifically, we describe a conversational agent that has the ability to exhibit bias towards one perspective or another as well as the ability to exhibit social-emotional behaviors that are designed to build solidarity. We describe a study in which we independently manipulate the social capability and goal alignment of the agent in order to investigate the impact on student learning outcomes, interactions, and perceptions. We observe a significant interaction effect between the social and goal alignment manipulation which suggests that the two strategies need to be considered together when designing tutoring systems. In addition, while there are less perceived differences in the student questionnaire data, we observe significant differences in student conversational behaviors in different experimental conditions. We suggest that an appropriate amount of tutor social behaviors can help to engage students in the conversation, and aligning with student goals can improve students' learning. In the future, we will further investigate how to design the tutor's social level and how to align with the learning objectives of both student partners in the conversation.

References

1. Strijbos, J. W. The effect of roles on computer supported collaborative learning, Open Universiteit Nederland, Heerlen, The Netherlands. (2004)
2. Baker, M., and Lund, K. Promoting reflective interactions in a CSCL environment. Journal of Computer Assisted Learning, 13, 175-193. (1997)
3. Kollar, I., Fischer, F., and Hesse, F. W. Computer-supported cooperation scripts - a conceptual analysis. Educational Psychology Review. (2006)
4. Weinberger, A. & Fischer, F. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. Computers & Education, Volume 46, Issue 1. (2006)
5. Paul, M., and Girju, R. Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models Export. In Proceedings of EMNLP. (2009)
6. Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., and Robinson, A. Tutorial Dialogue as Adaptive Collaborative Learning Support. Proceedings of Artificial Intelligence in Education. (2007)
7. Chaudhuri, S., Kumar, R., Joshi, M., Terrell, E., Higgs, F., Aleven, V., Rosé, C. P. (2008). It's Not Easy Being Green: Supporting Collaborative "Green Design" Learning, in Proceedings of Intelligent Tutoring Systems. (2008)
8. Isbister, K., Nakanishi, H., Ishida T., and Nass, C. Helper Agent: Designing an Assistant for Human-Human Interaction in a Virtual Meeting Space, CHI. (2000)
9. Morkes, J., Kernal, H.K. and Nass, C. Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of SRCT theory. Human-Computer Interaction, 14(4). (1999)
10. Wang, N. and Johnson, L. The Politeness Effect in an intelligent foreign language tutoring system. Intelligent Tutoring Systems. (2008)
11. Higashinaka, R., Dohsaka, K., and Isozaki, H. Effects of Self-Disclosure and Empathy in Human-Computer Dialogue, In: Proceedings of 2008 IEEE Workshop on Spoken Language Technology. (2008)
12. Brennan, S. E. and Clark, H. H. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition, 22. (1996)
13. Reeves, B., and Nass, C. I. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. (1996)
14. Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., and Nass, C. Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. Proceedings of CHI conference on human factors in computing systems. (2006)
15. Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Nass, C. Syntactic alignment between computers and people: The role of belief about mental states. Proceeding of the 25th Annual Conference of the Cognitive Science Society. (2003)
16. Concert Chat. http://www.ipsi.fraunhofer.de/concert/ (2006)
17. Blei, D. Ng, A., and Jordan, M. Latent dirichlet allocation. Journal of Machine Learning Research, 3. (2003)
18. Yano, T., Cohen, W., and Smith, N. Predicting response to political blog posts with topic models. In The 7th Conference of NAACL. (2009)

19 Paul, M. and R. Girju. Topic modeling of research An interdisciplinary perspective. In Proceedings of the the International Conference on Recent Advances in Natural Language Processing (RANLP). (2009)