Assessing the Participants in CSCL Chat Conversations

Costin Chiru, Traian Rebedea, Stefan Trausan-Matu, "Politehnica" University of Bucharest, Department of Computer Science and Engineering, 313 Splaiul Independetei, Bucharest, Romania
Emails: costin.chiru@cs.pub.ro, traian.rebedea@cs.pub.ro, stefan.trausan@cs.pub.ro

Abstract: In this paper, we present an application that can be used for assessing the contributions of participants to multiple chat conversations (that debate the same topics) according to different criteria, along with the ranking of the conversations considering a list of important concepts to be debated.

Introduction

Chat is one of the favorite environments for Computer Supported Collaborative Learning (CSCL) tasks that require online and synchronous textual interactions among participants (Stahl, 2006). Most of the existing tools for supporting chats are only aiming at facilitating the conversation without offering analysis instruments. One exception is PolyCAFe (Rebedea et al., 2011), a system which analyzes each user contribution and provides abstraction, visualization and feedback services for supporting learners and tutors. In another context, Chiru et al. (2011) start from the idea of topics rhythmicity in the participants' utterances in order to evaluate the chat quality and the personal involvement of each participant.

Still, what we consider to be missing in these systems is the lack of inter-chat comparison, especially when they are discussing the same topics. Thus, there is no easy way to assess all the chats conversations or their participants according to some predefined criteria imposed by the task at hand. To achieve this, we started from the heuristics proposed by Chiru et al. (2011) and developed two different evaluation methods – an intrinsic one considering only one conversation at a moment, and a multi-chat evaluation, that considers the conversation in the context of the whole corpus from which it has been extracted. We consider the importance of repetition expressed by Tannen (1989), not merely a repetition of words, but rather of lexical chains as we consider that all the words in a lexical chain refer to the same concept. Starting from repetition and other qualitative measures for participation in chats, we have proposed a set of heuristics used to assess the contribution of each participant.

Description of the Automatic Assessment

We developed an application for assessing chat conversations of undergraduate students at a Human-Computer-Interaction (HCI) class that were asked to debate about different web-collaboration technologies (forums, blogs, chats, wikis, etc.) in small groups of 5 students. For validating the results of the system, we have used a corpus consisting of 7 conversations ranging from 248 to 524 utterances per conversation, for a total of 2514 utterances. The same data has been used in validating previous systems (Rebedea et al., 2011).

In order to be able to assess the contribution of each participant to the conversation, we started from the heuristics proposed by Chiru et al. (2011), but we also investigated two other heuristics – *participant's knowledge* and *participant's innovation*. We computed *participant's knowledge* as the percent of the concepts introduced by the participant that were semantically connected with the ones imposed for debating (chat, blog, forum, wiki). In order to discriminate them from off-topic content, we used lexical chains built using WordNet (http://wordnet.princeton.edu/). Since these concepts were very specific to the HCI domain, most of the terms do not appear in WordNet with the required senses. Thus, we had to develop a taxonomy of concepts related to each of the conversation topics and we found 76 concepts related to *chat*, 44 related to *blog*, 63 terms for *forum* and 31 for *wiki*.

The second introduced heuristic, *participant's innovation*, was computed as the number of concepts introduced in the conversation by each participant and represented the degree of new information introduced by that person. The computed values for each heuristic have been normalized with respect to minimum and maximum values obtained for each of them. For this task we had two alternatives: one related to a *single chat* and another considering *all the chats* that were used. The first alternative can be used for evaluating the contribution of each participant in a single conversation, while the second can be used for evaluating a person considering also the activity of all the other persons involved in similar conversations from the same corpus.

After that, we combined the quantitative heuristics proposed by Chiru et al. (2011) - number of replies, activity, absence, persistence and repetition - considering that they all characterize the involvement of a participant. Therefore, we computed involvement as the sum of the values obtained for each of the 5 considered heuristics. The final score for each participant was obtained as an average value of *involvement*, *knowledge* and *innovation*.

The application also compares the conversations one to each other in order to decide which one of them has achieved the best results considering the fact that the discussion topics were externally imposed by the tutors.

Analysis and Results

In order to evaluate the quality of the assessment, we asked 4 HCI experts to manually assess the conversations that were part of our corpus (Rebedea et al., 2011). We ended up with 15 reviews: 4 for chat 117, 3 for chat 116, 2 for chats 118, 119 and 120 and a single review for chats 125 and 126. At the same time, we asked the chat participants to rank their colleagues with respect to their activity in the chat. As a third way to evaluate the results of the proposed system and assessment factors, we also considered the scores provided by PolyCAFe.

The overall correlation was lower than expected, having an average correlation of 0.60 with the tutors, 0.66 with PolyCAFe and 0.47 with the students for the single chat analysis and even worse for the newly-proposed multi-chat evaluation – ranging between 0.24 and 0.37. Moreover, we observed that very large values for the standard deviation have been obtained showing that the overall score is not very suitable for assessing the contribution of the participants. These results are also poor when compared to the correlation between tutors-students (average correlation $r = 0.87$, $\sigma = 0.19$), tutors-PolyCAFe ($r = 0.94$, $\sigma = 0.05$) and students-PolyCAFe ($r = 0.85$, $\sigma = 0.16$).

Therefore, we have tried to see which one of the three components of the overall score is responsible for these problems. We have computed the correlation between the scores provided by tutors, students and PolyCAFe with our results for each of the three components: involvement, knowledge and innovation. The correlation with the involvement heuristic proved to be extremely good, $r = 0.90$ with tutors and PolyCAFe and $r = 0.83$ with students in the case of single chat analysis and 0.86, 0.87 and 0.81 in the case of multi-chat analysis. The standard deviations are also good as they have very low values (between 0.10 and 0.16). The innovation heuristic also seemed to be extremely well correlated with the gold standard, obtaining an average correlation of 0.90 with tutors, 0.93 with PolyCAFe and 0.86 with students with $\sigma \leq 0.10$. These results were the same in both cases of the single and multi-chat analysis.

Finally, the correlation between the gold standard and the knowledge heuristic provided us a great surprise: most of the time, our results proved to be anti-correlated with the gold standard. For this heuristic, both the single and multi-chat analysis had the same results: the average was -0.60 with tutors, -0.57 with PolyCAFe and -0.67 with students, while the standard deviation was 0.22.

Considering this strong anti-correlation for this factor, the main problem with our assessment method proved to be the way we considered knowledge in the final score. Moreover, it seems that our current method for assessing the knowledge of a participant was less correlated with the golden standard even if we would have considered the absolute value of the correlation value ($r = 0.60$ compared to $r = 0.90$ for the other two heuristics).

Encouraged by the above finding, we continued with the 5 factors influencing the involvement heuristic in order to identify which are the most important ones and which can be ignored. Starting from the obtained results, we identified another heuristic that was anti-correlated with the gold standard: *Persistence*. Besides this heuristic, we also identified that *Activity* is not well correlated with the manual annotation ($r = 0.37$ $\sigma = 0.51$). In conclusion, from the initial 5 heuristics considered together to characterize the participants' involvement, only 3 actually provide important evidence to motivate their use: Number of replies, Absence and Repetition.

Conclusions

We have shown that the overall score computed by the application is not very reliable especially when compared to other systems. However, when analyzing each component used to compute the overall score, we have found that some of the heuristics perform quite well and that the overall results are affected by only a single factor. The heuristics that proved to work best are innovation and involvement, while the one used for assessing the knowledge of the participants was either poorly designed or poorly interpreted. Maybe the most important result in this paper is the methodology of how to identify which heuristics work best and which are the ones that should be avoided if a combined score should be computed by a given application for assessing CSCL chats or other learning tasks.

References

Chiru, C., Cojocaru, V., Trausan-Matu, S., Rebedea, T. & Mihaila, D. (2011). Repetition and Rhythmicity Based Assessment for Chat Conversation. *ISMIS 2011, LNCS 6804*, pp 513-523.

Rebedea, T., Dascalu, M., Trausan-Matu, S., Armitt, G. & Chiru, C. (2011). Automatic Assessment of Collaborative Chat Conversations with PolyCAFe. *EC-TEL 2011, LNCS 6964*, pp. 299-312.

Stahl, G. (2006). *Group cognition. Computer support for building collaborative knowledge*. Cambridge: MIT Press.

Tannen, D. (1989). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*: Cambridge University Press.