A Computation Model for Quantitative Conversation Analysis in Collaborative Learning Settings

Abstract

In this work, we aim to identify quantitative metrics that can be used to predict task success or discriminate between successful and unsuccessful groups involved in a collaborative learning task, using text based chat transcripts. We draw from theories of argumentative knowledge construction and systemic functional linguistics to propose method for conversation analysis that can be an automated using machine learning techniques. Our results show striking differences between conversation trends that can be used to classify groups as successful or not.

Keywords. Task based dialogue, knowledge sharing, automated discourse analysis, Computer Supported Collaborative Learning

1 Introduction

Interest in automated Discourse analysis has been growing over the past few years . Initial work in this field looked at simple measures involving quantitative information about levels of participation [1]. However, such simple measures do not usually have much predictive power for task success [3]. Therefore, content analysis became a popular technique to analyze information in transcripts of asynchronous groups. Researchers have used chat transcripts to investigate the process of social construction of knowledge [7] or critical thinking (Bullen,1997; Newman, Webb & Cochrane,1995) to differentiate from earlier methods that used surveys, interviews case studies, statistical measurements etc. which did not shed much light on the quality of learning taking place. In general, the aim has been to reveal information that is not situated at the surface of the transcript. [2] provides a fairly comprehensive listing of some of the initial work in this area that focuses on content analysis. A majority of these methods however, focus only on the qualitative aspect of analysis and de-facto, require manual involvement. While they provide fascinating insights into the nature of collaborative conversation, automating such analysis and mapping them to quantitative metrics has inherent value, especially when trying to deal with the kind of scale we see today, with the prospect of MOOCs replacing traditional classrooms, thus making manual analysis impractical. This work proposes one such scheme that draws from the concepts of argumentative knowledge building [4], systemic functional linguistics [8] and the theory of active learning [9], to analyze collaborative conversation in a way that can be automated using current machine learning techniques. Another interesting possibility with this technique is it's ability to capture global context in a conversation to provide a sense of the overall trajectory of a conversation, which can be used to motivate tutor interventions based on global phenomena rather than just local context triggers.

2 Related Work

As mentioned before, [2] contains a fairly comprehensive survey of initial conversation analysis methods. These include assessing cognitive and meta cognitive knowledge [1], critical thinking, social constructivism in knowledge construction (Gunawardena et. al, 1997), social network theory for analyzing interactional exchange patterns (Fahy et. al ,2001)[4] etc. Some of them still rely on measuring participation in some form or the other.
A number of frameworks have been proposed for analysis of discourse. In [4], a framework based on argumentative knowledge construction is proposed that considers multiple dimensions of analysis:

1. Participation,
2. Heterogeneity of Participation,
3. Epistemic dimension,
4. Argumentative dimension and
5. Social Modes of Reconstruction.

There are also theories that investigate the efficacy of different argumentation methods (differentiating between single arguments and argument sequences [5]) and consensus building techniques (weighing integration oriented consensus against conflict oriented consensus building based on transactivity [6]).
Apart from such qualitative methods of analysis mentioned, some quantitative metrics for predicting task success have received attention in recent years. One example includes [n-8] which looks at priming ratio and matched word ratio among users to identify convergence in task oriented dialogue, which has been known to predict dialogue learning outcomes. However, the focus here is on capturing convergence between the same two interlocutors over an extended period of time.
In [n-9], two LSA based approaches are proposed that use semantic similarities between groups known to have performed well, and the group that is being assessed, to predict success. In the first approach, they directly use LSA estimated performance scores to predict performance. In the second approach, they look at frequency of tags associated with each individual turn in conversation, to predict success. This is reminiscent of the initial works on measuring level of participation. But, it goes a step further by assigning a role in conversation to each turn. However this study looks only for the presence of relevant discourse without worrying about importance of sequential information and context in conversation. In a setting like ours (described later), presence of an tutor system that regularly intervenes and a set of clearly defined short term objectives being regularly introduced during the group activity, group members tend to stay focused and on topic.
Importance of sequential (context) and structural (reply structure, segmenting of conversation) information has been explored in prior work. [n-10] uses Hidden Markov Models (HMM) to identify sequential characteristics in the conversations of successful and unsuccessful groups by assigning role based tags to individual turns and looking at the context in which each of these turns occur. [14] further analyzes the HMM models by identifying the most characteristic conversation sequences of both these types of groups. The identified features were good enough for an overall classification task. However, membership probabilities with respect to trained HMM models needn't directly correspond to how well a group performed. Moreover, in multi-threaded conversations, multiple interlocutors could lead to overlap between different parallel sequences which is not handled well by an HMM.

Another form of sequence modeling is through the use of Systemic Functional Linguistics [8], through the identification of 'Martin Sequences' in conversation. In simple terms, a martin sequence consists of question answer pairs along with the following agreement, or alternatively, a new proposal with its corresponding agreement. Recent work [15] has shown that identification of such sequences in conversation is possible with a high degree of accuracy and can even be done in real time, using machine learning techniques. Our work uses this representation of a sequence to account for structure and context for each turn in conversation, while determining its role in conversation and deciding how it contributes to the discussion.

Given the nature of our dataset (described later), we are primarily concerned with identification of knowledge sharing in discussion. This aspect of conversation has been studied before using machine learning methods [13], statistical discourse analysis [16], uptake graphs [17] etc. This prior work is focused specifically on the process of group problem solving in collaborative learning, which is what we are interested in. [18] attempts a similar task of identifying cultural differences by looking at variation of a single dimension of authority levels [19] of participants over time, but finds that simply monitoring these trends was not predictive of success. However, this work neglects the sequential nature of conversation and dependency on context for deciding roles of turns in conversation. In our work, we hypothesize that accounting for this structure and context in conversation to decide roles and monitoring how they affect two parameters, namely consensus and confusion could yield trends in conversation predictive of task performance. Moreover, by looking at the aggregate of consensus gained and confusion reduced, we could compare groups based on these parameters, which could now serve as quantitative metrics.

3 The Collaborative Task

In our work, we analyze a dataset that consists of chat transcripts of college freshmen taking a chemistry course. Prior to the collaborative session, a pre-test was conducted to assess the initial levels of knowledge that each of them possessed. After this, students were randomly assigned into groups of 3-4 people and each group member was given some specific knowledge in a Jigsaw fashion [20]. Students collaborated for 75 minutes in an online chat setting, with an intelligent tutor system periodically introducing new questions or task objectives and occasionally intervening to re-voice a statement or target specific group members. During the course of these 75 minutes, students read their individual training material (unique for each student in a group) and collaborated over a shared assignment steered and directed by the automated tutor. After the session, each student had to take a post test. The change in pre to post test score is what we use a metric for learning outcome.

For the purpose of our analysis, we looked at the top 3 teams and the bottom 3 teams out of a total of 8 teams that consented to have their scores used for research purposes. We chose the top and bottom 3 so as to obtain a clear distinction between successful and unsuccessful teams. In order to decide the ranking of teams we run a linear regression that predicts the post test score using the pre test score for each student. We find the residuals for each student. For each group, we find it's residual as the average of the residuals of its members. This approach of averaging residuals looks at how much better than expected students belonging to a particular did.

The combined size of the chat transcripts is about 4500 lines of conversation. These have been manually tagged based on different annotation schemes (described next) for the purpose of our analysis.

4 Annotation Schemes

We use two annotation schemes :

1. The Negotiation Scheme
2. Active Learning Scheme

4.1 The negotiation scheme

It is based on the Systemic Functional Linguistics framework [8] and annotation based on this gives us the building blocks with which to study interaction in conversation. In our analysis scheme, our concept of sequences in conversation and sequence modeling of context stems directly from this representation. This analysis is also used in [18].
This scheme differentiates between the following 5 types of moves :

1. K Type Moves:
   (a) K1 : Represents the giving of knowledge. It is usually a statement of fact or opinion.
   (b) K2 : It involves asking for someone else's knowledge or opinion.
2. A Type Moves :
   (a) A2 : Involves providing instructions or dictating action.
   (b) A1 : Involves narration of action, acceptance of directives for action etc.
3. Follow Up moves :
   These usually involve acknowledgement of an A1 or a K1 move. It does not contribute any additional information.
4. O moves :
   These moves include floor grabbing moves, false starts, preparatory moves and any other non-contentful contributions.

In our work, annotation based on this scheme was done by two annotators independently and a kappa of 0.76 was achieved.

4.2 Active Learning Tags

This annotation is influenced by the generalized active learning type roles proposed in [9]. We adopt a simplified version of that breakdown of conversation acts by considering just 5 types of moves. These are further justified by the social modes of co-construction as highlighted in [n-2]. A similar coding scheme is also used in [21]. However, a key difference is that in that work, classification into the following categories is based on a sentence opener model, whereas in our work, it is based on the position of a move in its sequence. The 5 categories are :

1. Proposal : These moves introduce a new concept or idea. As per [4], this would correspond to Externalization type moves. They are usually K1 type moves and occur at the beginning of a sequence.

2. Doubt : These moves target proposals and question them. They are K2 moves.

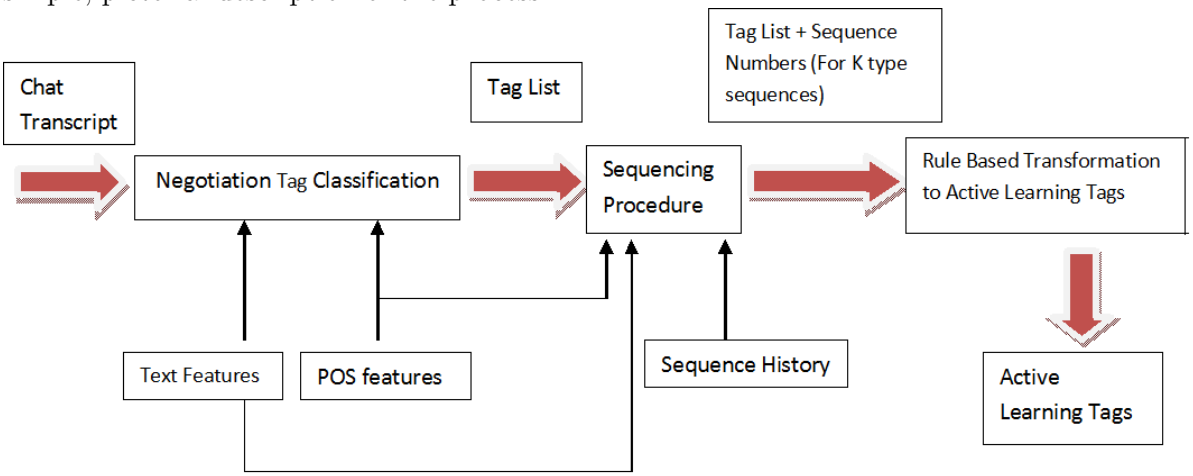
3. Clarification : These are replies to doubts. They are usually K1 moves that occur after a K2 in the same sequence.

4. Agreement : These indicate consensus between a K1 speaker on the same sequence and either another person or a K2 speaker on the same thread.

5. Comment : These are arbitrary statements. O moves are comments. Off topic statements, floor grabbing moves, pauses etc. are generally classified as comments.

5 Process Pipeline

In this work, we propose a method for conversation analysis which uses the change in authority levels [19], within a sequence structure [8] [15], to classify each turn in a conversation into active learning [9] based tags as defined in the annotation schemes section. Below is a simple, pictorial description of the process :



The motivation for using this pipeline instead of using a direct classifier to generate active learning tags is described in the results section.

6 Transformational Rules

- If a K1 starts a sequence/is first K type move in a sequence, then it is a pr. Subsequent K1 moves from same speaker in same sequence are comments, but the first K1 by any other speaker in the same sequence is labeled as a clarification.

- However, if the same speaker gives a K1 after an f move by someone else directed at his original K1, then the first such K1 is a clarification, and all following K1s by him are comments.

- K2 is always a doubt.

- In response to a K2 if we get an f/K1 (need not be immediately after, it can be separated by 'o' moves, from the K2) from a different speaker then it is put as a clarification. Any further f/k1 moves from those speakers in the same sequence are put as comments.

- An F move from the K2 asker is put as an agreement if it occurs after at least one K1 from somebody else, otherwise it is a comment.
- O moves are put as comments.
- If in response to a proposal (K1), a different speaker gives an f move, it is put as an agreement. The very next K1 (but NOT an f move) from this other speaker is put as a cl. Any further K1s from him are put as comments. Any further f moves from this other speaker will be put as comments.

7 Representation of Interaction in Dialogue

In our analysis of dialogue, we follow an approach very similar to [n-18] wherein we assign weights to each category of active learning tag, for two parameters : Consensus & Confusion. *Consensus* value indicates the level of agreement between group members. If consensus increases, group members are agreeing, if it reduces, there is some unresolved conflict/disagreement. *Confusion* indicates to what extent questions are being answered, or positions are being defended or clarified. A reduction in confusion indicates prompt replies to doubts or challenges, or clarifications of a proposal by peers. Our reasons for choosing these two parameters and setting the weights towards these parameters for each active learning tag in conversation, stem from two of the dimensions of analysis outlined in [4] :

1. Argument Dimension
2. Social modes of Consensus Building

7.1 Argument Dimension

This looks at how learners in a collaborative setting construct and balance arguments supporting their knowledge or hypotheses and defend them. Arguments themselves are of two types : Single arguments, which are self sufficient in that they contain a claim, along with the grounds that support it as well as qualifiers that outline the specific cases in which it holds. However, studies show that even adult learners tend not to construct such arguments on their own (Kuhn,1991,[23]).
Much more common is an argument sequence, consisting of arguments, counter arguments and replies. In our model, this would manifest itself as a proposal, followed by a doubt and it's clarification, or alternatively, a proposal, followed by another proposal (contra-proposal) and it's agreement. With the construction of such argumentation sequences, members of the collaborating group acquire multiple perspectives on a problem and this enables them to solve future problems [24].
So we see that for an effective argument sequence, we should ideally have a buildup of consensus, after deliberation (which should reduce confusion). Essentially, increase in consensus indicates agreement on a proposal, while decrease in confusion indicates how well deserved this consensus is.

7.2 Social modes of Consensus Building

The following are three modes of consensus building :

1. Quick Consensus Building:
   This involves accepting a new proposal by a team mate, in order to move on with the task. This type of consensus does not indicate consensus accrued over a long series of deliberation. Rather, it serves as a coordination function (Clark & Brennan,1991). In our model, such quick consensus building would manifest as a pattern of a proposal followed immediately by an agreement in the same sequence. We thus set consensus weights for proposals and agreements to indicate a build in consensus, without a reduction in confusion to qualify it, whenever such a sequence occurs.

2. Integration Oriented Consensus Building:
   This is characterized by the views/proposals of one individual being taken up by his peers. Learners may modify or give up initial beliefs and correct themselves based on views of their peers. However, studies to date have been inconclusive about whether this type of consensus is really indicative of better individual knowledge acquisition. Moreover, this type is consensus building is rarer than other forms.

3. Conflict Oriented consensus building:
   It involves proposals or ideas being challenged with specific problems or instances, which require learners to try multiple perspectives and defend their positions on an issue. In [6], a factor of transactivity was identified, which was found to relate positively with learning and knowledge acquisition. Conflict oriented consensus building has the highest transactivity out of all the social modes outlined in [4]. We thus set the weights of our active learning tags in such a way as to incentivize conflict oriented consensus. As a gold standard we consider a set of moves comprising a proposal, followed by a doubt, it's corresponding clarification, and the consequent agreement to the proposal as a good manifestation of conflict oriented consensus.

In summary, given the positive and negative conversation practices identified above, in our model we must ensure:

1. Quick agreements are not qualified by a decrease in confusion. Thus, proposal agreement pairs alone do not build confusion.

2. A dictatorial style of continuous proposal must be discouraged as it does not involve building argument sequences. So lone proposals must not be associated with building consensus or decreasing confusion.

3. Each contra-proposal/doubt must be qualified by a reply or answer, as required in argumentation sequence. Thus lone doubts must be associated with both an increase in confusion and a decrease in consensus. The reply/clarification to this must negate both of these effects while decreasing the confusion a little further as the doubt/clarification is an example of deliberation in argument sequencing.

4. In order for conflict oriented consensus building to be achieved, an agreement to the reply is necessary. Thus consensus is not built just by a doubt clarification pair.

Similar findings are also reported in [14] which used Hidden Markov Models to identify sequential patterns/trends in actual conversations which were used to classify groups as good or bad. A clustering approach (augmented by multi dimensional scaling to account for large number of outlier points in data) was then applied to identify conversational sequences that contributed to differentiation of good groups and bad groups. Below is the summary of the results presented in that work:

Table 1: General Patterns for each Effective Group

Group A_e	**Group** B_e	**Group** C_e
Receiver requests information about something (K2/doubt)	Receiver requests information (K2/doubt)	Sharer explains or illustrates something (K1/proposal)
Sharer explains or illustrates something (K1/proposal)	Sharer provides explanation (K1/clarification)	Receiver motivates / encourages(F/agreement)
Receiver agrees (F/agreement)	Receiver requests further clarification (K2/doubt)	
	Sharer provides further clarification (K1/clarification)	

Table 2: General Pattern for each Ineffective Group

Group A_i	**Group** B_i	**Group** C_i	**Group** D_i
Sharer proposes something (K1/proposal)	Sharer attempts to explain something (K1/proposal)	(F/agreement)1. Sharer proposes something	Receiver requests explanation of KE (K2/doubt)
Sharer explains or gives instructions for action (K1/proposal contd.)	Receiver acknowledges	Receiver doubts (K2/doubt (unresolved))	Sharer explains poorly (O or lack of K1) (no further discussion)
Receiver acknowledges (F/agreement)			

Thus we find that these results of the analysis of actual data, as reported in [14] seem to validate the positive and negative conversational trends identified in [4]. Using these trends as guidelines we have set the weights for each active learning tag towards consensus and confusion to incentivize good sequences in conversation and penalize bad ones.

Active Learning Tag	Confusion	Consensus
Proposal	0.3	-0.3
Doubt	1	-0.5
Clarification	-1.3	0.5
Agreement	-0.3	0.8
Comment	0	0

As an example consider the following sequences:

Sequence	**Trend**	**Consensus Change**	**Confusion Change**
Proposal → Agreement	Quick Consensus	+0.5	0
Proposal →Clarification→ Agreement	Integration Oriented Consensus	+0.5	-1.3
Doubt → Clarification (No Agreement)	Argument Sequencing	0	-0.3
Doubt → Clarification → Agreement	Argument Sequencing	+0.8	+0.5
Proposal → Doubt → Clarification→Agreement	Conflict Oriented Consensus Building	+0.5	-0.3

8 Trajectory Based Approach

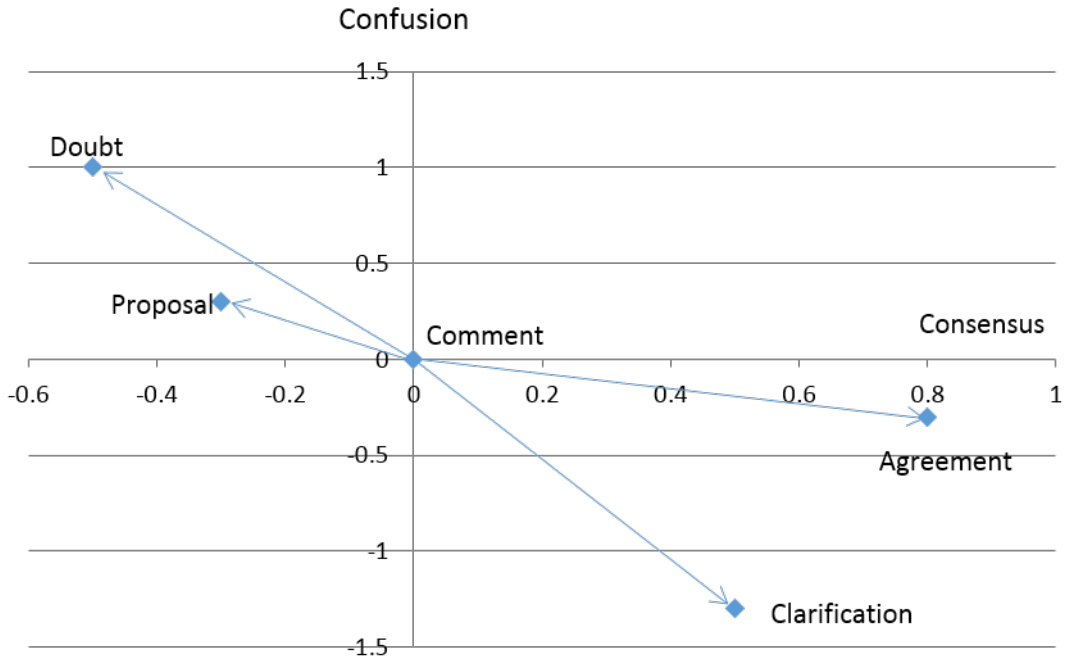
For the purpose of our analysis, we are interested in assessing knowledge exchange during a collaborative learning session. Moreover, we have identified that the correct level of granularity for our analysis is not at the individual turn level or the utterance level, but rather, the sequence level. This is motivated by the following reasons:

1. The sequence level captures the relevant context (sequential information) for each utterance it contains, which can then be used to identify the active learning role played by it, using rule based transformations.

2. The sequence level brings together related contributions (based on content) by the collaborating group members, which are usually related by a causal/reply structure.

3. The existence of appropriate context and causal/reply structure makes it ideal for identifying the trends highlighted in the previous section, which have been shown to be associated with knowledge acquisition.

Keeping this in mind we extract those sequences identified during the pipeline process, which contain only K type (knowledge based) moves. We then generate active learning tags from them, using the rule based transformation. Once active learning tags are obtained, the analysis involves tracking the variations in consensus and confusion, given the weights associated with each of these tags. This is done using a 2-dimensional graph.
However, we also need some concept of time to capture the temporal/sequential aspect of conversation. In our analysis, each conversational move represents a unit step in time. Thus for each such unit step in time, a certain amount of consensus and confusion is added, depending on the type of conversational move that occurs next. Therefore, each active learning role/tag can be thought of as a vector in 2 dimensions, with one vector being added on for each unit of time. The figure below makes the concept clear.



The results of plotting graphs using this approach are discussed in the results section.

9 Results

9.1 Conversion Pipeline

For the process of converting raw chat data, to the active learning tags, the first approach was to build a Logistic Regression classifier using LightSide [22], a weka based machine learning software, specialized for text mining. The features used included unigrams, bigrams, POS bigrams (to capture some stylistic aspect of speaker turns) & binary n-grams. We removed stop words and included punctuations since some features such as question marks can be very helpful (for example, to classify doubt type moves). The resulting classifier had a Kappa of 0.425 with 57.7% accuracy, with the most problematic misclassification occurring between proposals and clarifications and vice-versa as these two moves contribute oppositely towards the two parameters of confusion and consensus. This is especially problematic since the weights towards confusion and consensus for these two conversational moves are almost opposite.

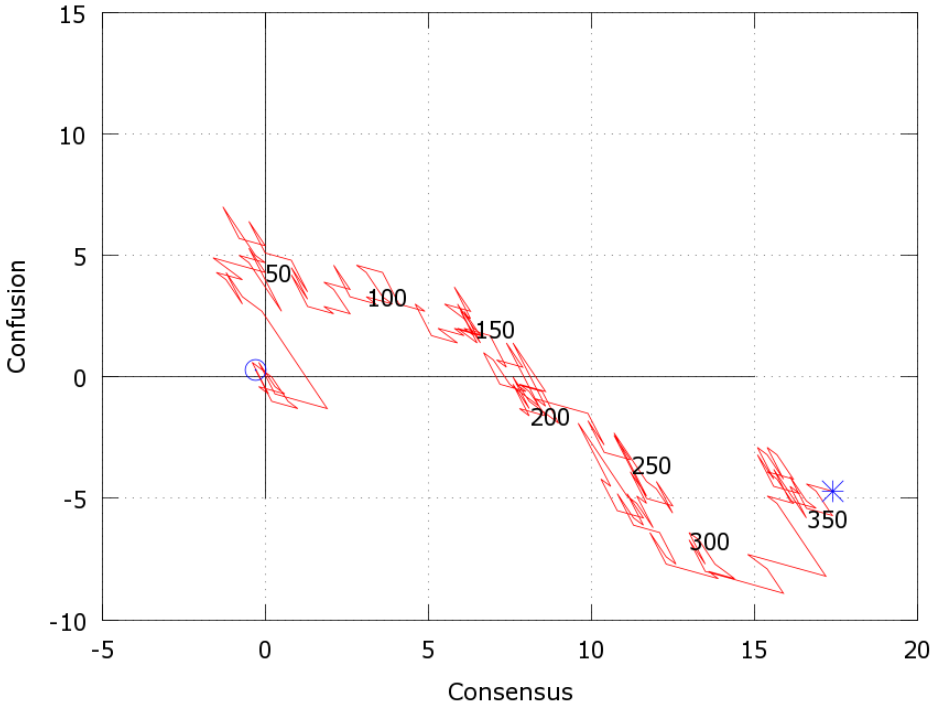
As a contrast to this, when the pipeline based approach was tried using fully human annotations for both the negotiation tags and the sequencing as an input to the rule based conversion in the last step, a kappa of 0.87 was achieved with respect to the human annotated active learning tags. This suggests a strong co-relation between the variation of authority levels within the sequence structure, and the active learning roles identified manually. In the results presented next, the active learning tags generated with this method have been used. Encouraged by this, a partially automated approach was attempted, with the first step of the pipeline (negotiation tag classification) being automated. A Logistic Regression classifier was built with LightSide, using the same features as before. This classification step had a kappa of 0.7022, with an accuracy of 76%. The improvement in kappa is primarily since K1 moves encompass both proposals and clarification, between which there were a large number of misclassifications in the previous model. When coupled with the manual sequencing as an input to the final step of the pipeline, the active learning tags thus generated matched

the human gold standard annotations, with a kappa of 0.53. This is higher than the kappa of 0.42 achieved with a direct classifier. However, it must be mentioned that the human sequencing might be sub-optimal with respect to the automatic negotiation tags generated. But, the kappa value suggests that even so, we do better using the pipeline than using a direct classifier.

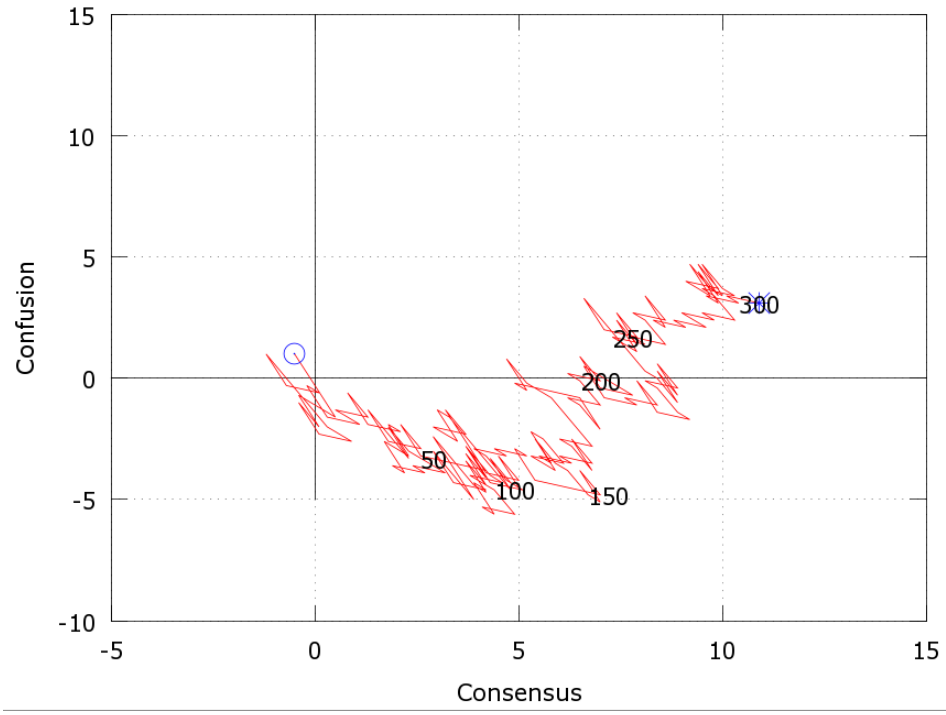
9.2 Trajectory Analysis

For each of the six groups considered in our dataset, active learning tags were generated. Using these tags, the consensus vs. confusion graphs were plotted. The following are the results:

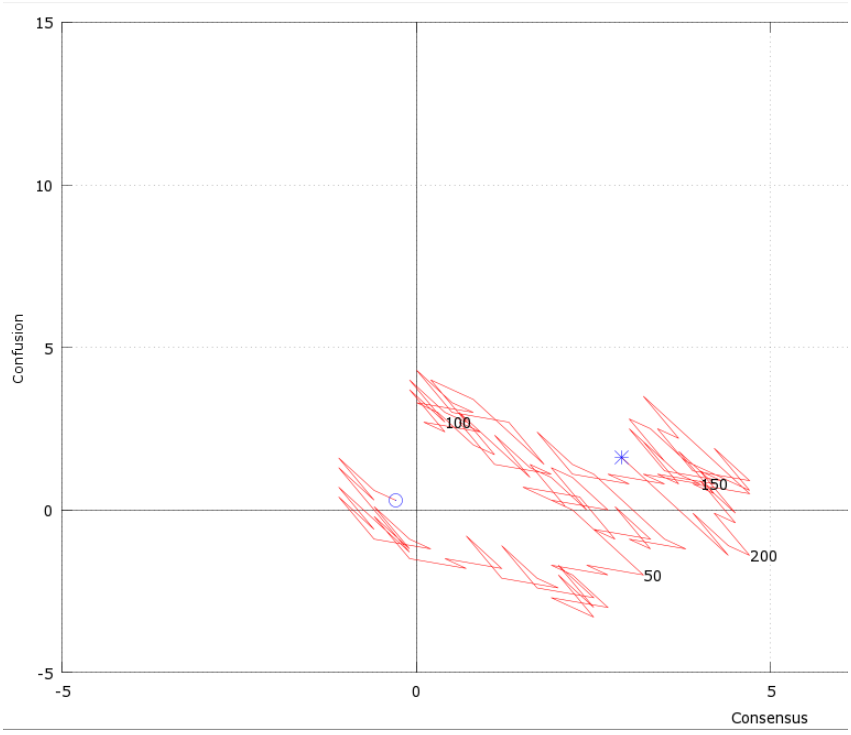
1. **Group 3**: Average Residual Score for pre test to post test gain of all team members (ARS) = +1.944



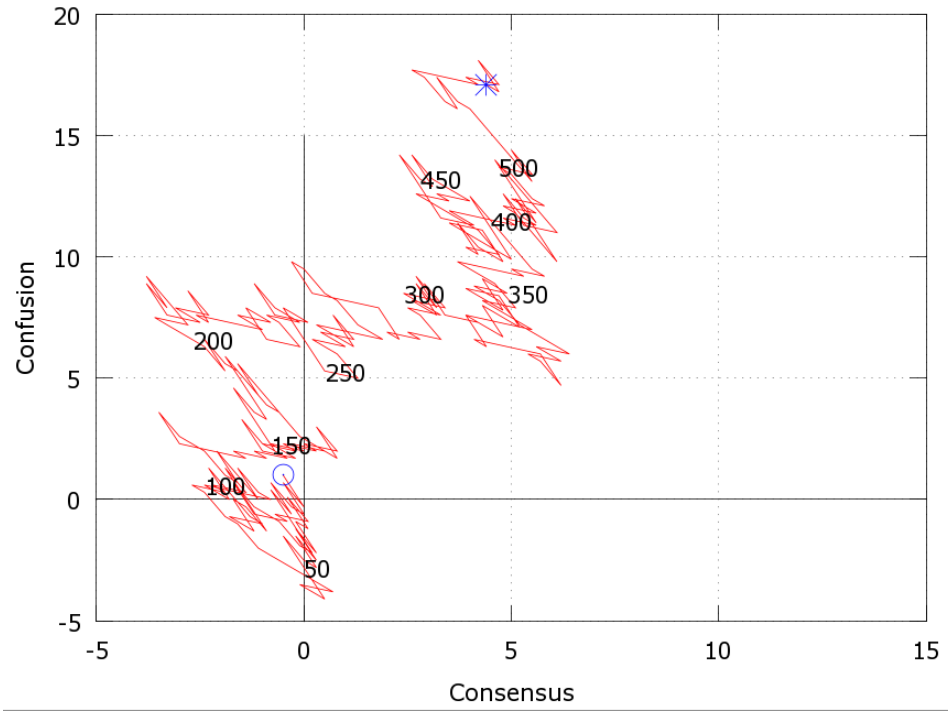
2. **Group 7**: ARS = +1.81737



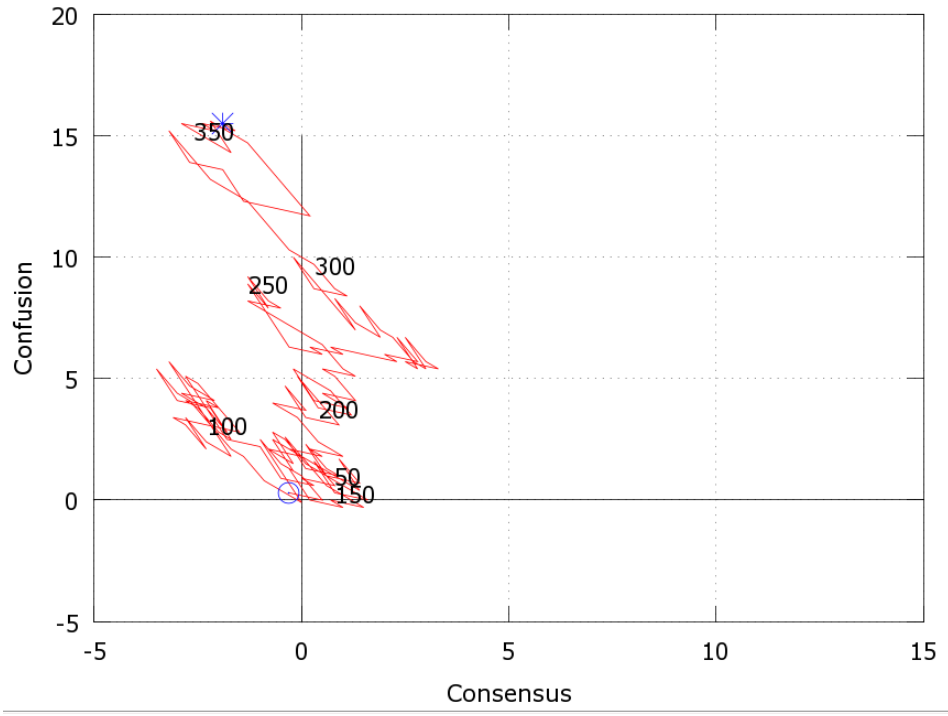
3. **Group 5**: ARS = +1.72384



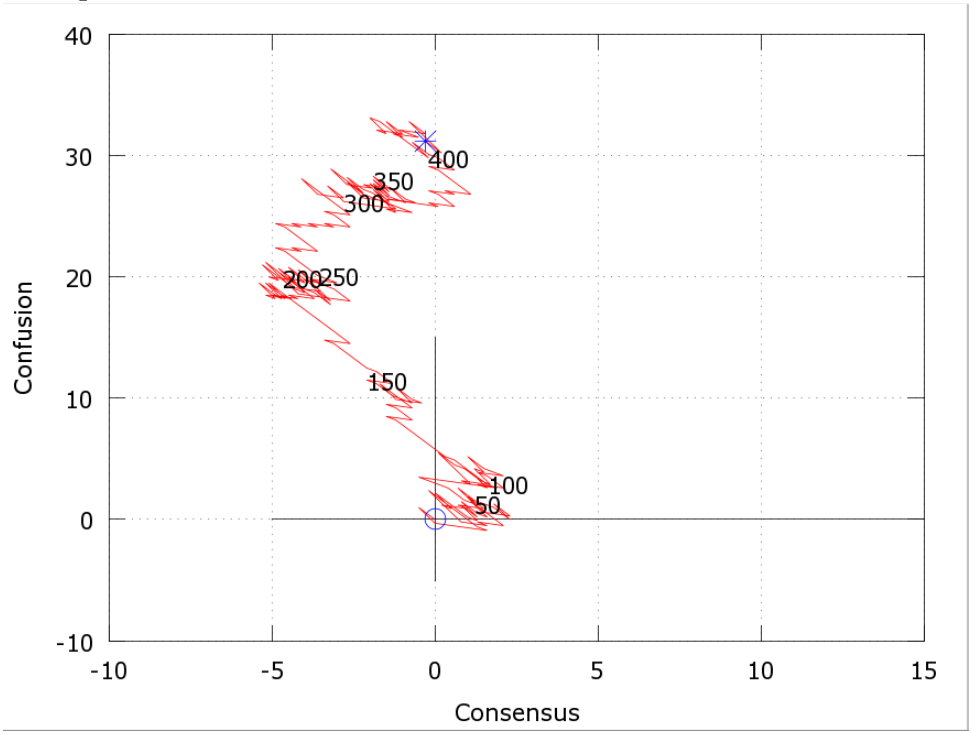
4. **Group 12**: ARS = -1.81915



5. **Group 9**: ARS = -1.84965



6. **Group 6**: ARS = -2.43217



9.2.1 Use of Average Residual Score (ARS)

To decide which teams were successful and which of them weren't, we use linear regression based mapping from pre-test scores to post test scores, for each student participating in the experiment. We then find the residuals for each student. This is an indication of how much better (or worse) than expected he performed. For each group, we then find the average of the residual scores of the group's participants and use this as a metric for overall group success.

By looking at the above graphs, the first observation is the drastic difference between the general orientation of the Consensus/Confusion graphs of the successful teams (high positive ARS) and the unsuccessful ones (high negative ARS). Successful teams tend to build up consensus while minimizing build-up or even decreasing confusion. The opposite holds true for unsuccessful teams.

It is also interesting to note that some teams such as team 12 managed to build consensus despite having built confusion. While this may sound counter-intuitive, the reason is because this team was characterized by group members who constantly made new proposals. Most of these were met by quick agreements without any argumentation. In our model, such proposal agreement pairs achieve consensus but are not qualified by a decrease in confusion. The build-up in confusion indicates that questions that did come up went unanswered.

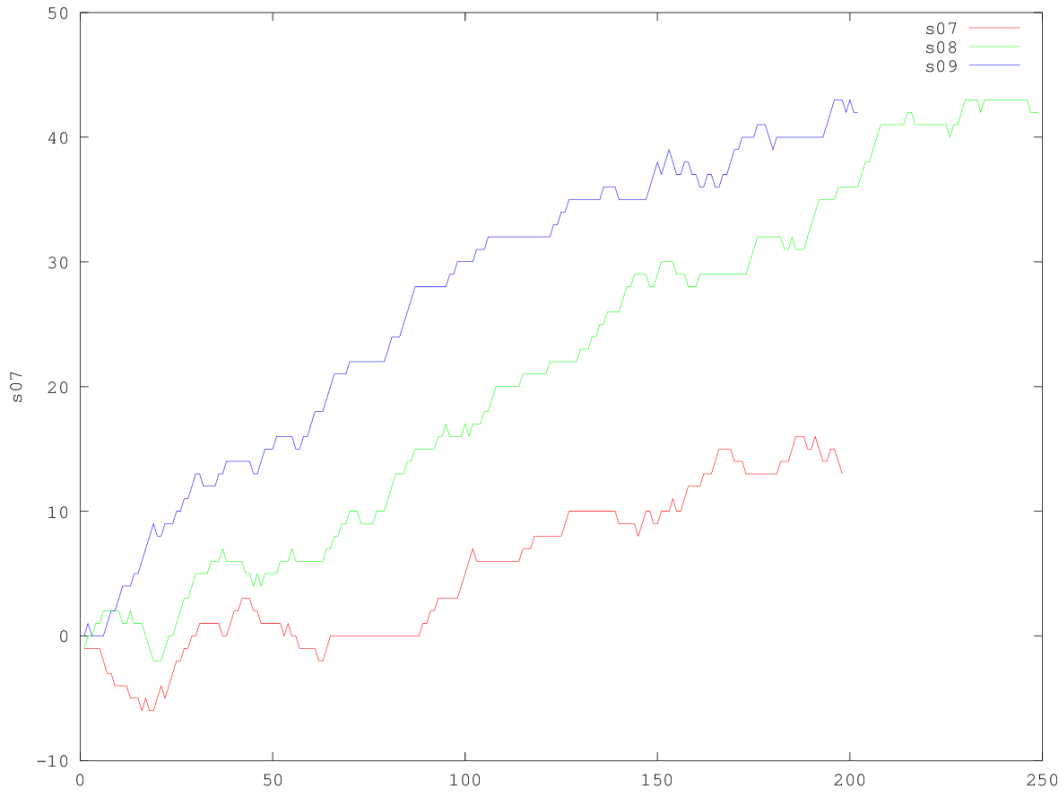
This is also evidenced by the authority graphs (shown next) where we find that 3 out of 4 group members in group 12 achieve a very high authority level, indicating that they had a proclivity for issuing only K1 type moves. The fact that 3 out of 4 members achieve a high authority score also shows that hardly any K2s occurred. As a result, most of the K1s were proposals and not clarifications. In group 6 on the other hand, we observe that from the 100th to the 250th conversational move, there is a steep increase in confusion. This is due to a large number of unresolved doubts cropping up. This trend could alert the tutor system to intervene and encourage clarifications or even provide clarification.

In group 7, we find a healthy trend in conversation up until around the 150 conversational moves. After this we find a trend of gradual but growing confusion. This was primarily due to the group members trying to rush through the final stages of the assignment, which led them to neglect each other's proposals. This is why the buildup in confusion is gradual and not steep (since increase in confusion associated with proposals is lesser than doubts).This trend could be used to stop the tutor from enforcing a rigid macro script that with a strict time limit.
Another interesting point to note is that length of discussion is not predictive of performance. Despite having restricted our analysis to only K-type sequences that are on-topic, most teams (with the exception of group 5) have conversations of similar length.
Another application of these graphs is to look at the flow of the overall conversation. In group 7, we notice that towards the end of the conversation, there is a steep increase in confusion. This could serve as a red flag to the tutor system, indicating that the group members have run into some trouble. In group 12, we notice that from the 200th to approximately the 300th line of conversation, there is an almost horizontal movement in the graph. This indicates a series of quick agreements. This could once again serve to alert the tutor

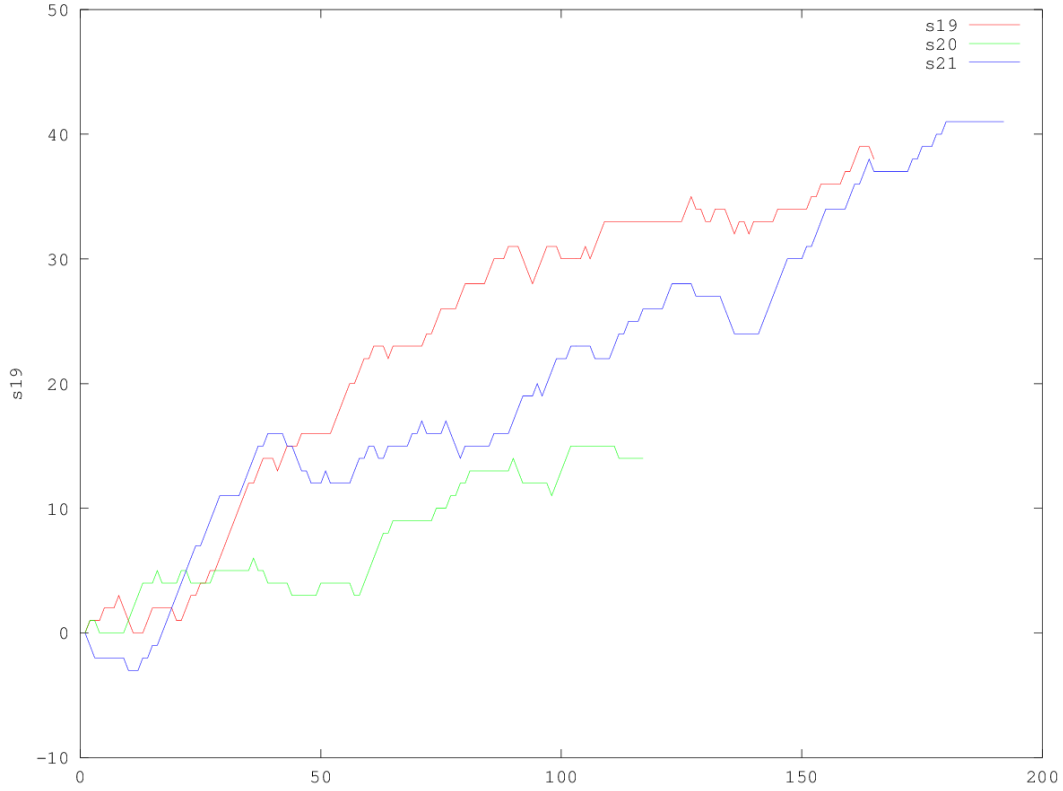
9.3 Authority Graphs

As a contrast to our method, the analysis based on authority graphs proposed in [18] was repeated for these groups. The graphs track the variation in K type authority level of each student of a group over the course of the entire conversation. Each K1 type move is associated with an increase in authority of +1 while a K2 has a weight of -1. The following are the results:

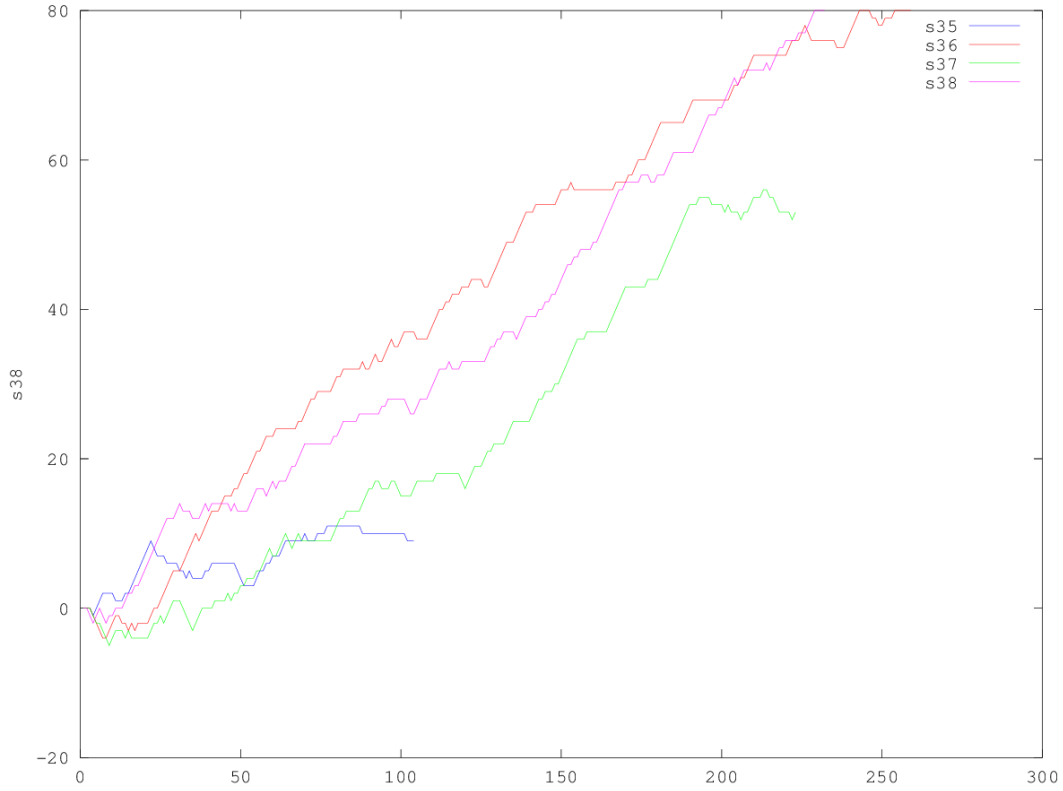
- **Group 3**: Average Residual Score for pre test to post test gain of all team members (ARS) = +1.944



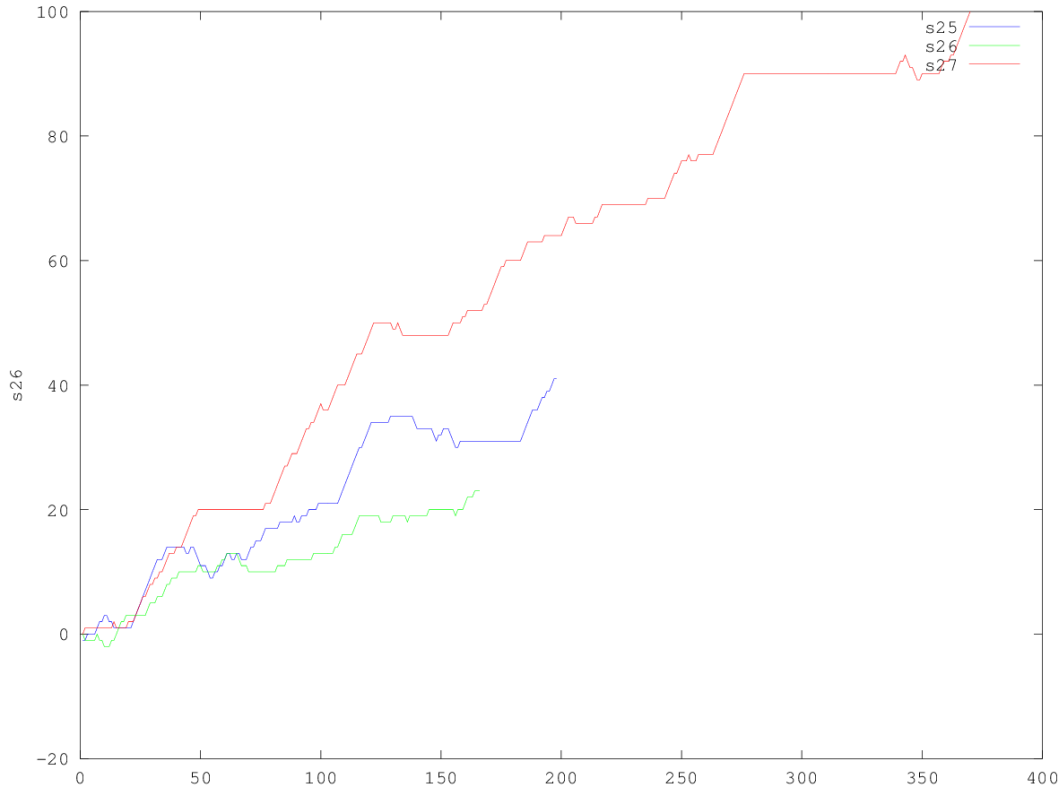
- **Group 7**: ARS= +1.817



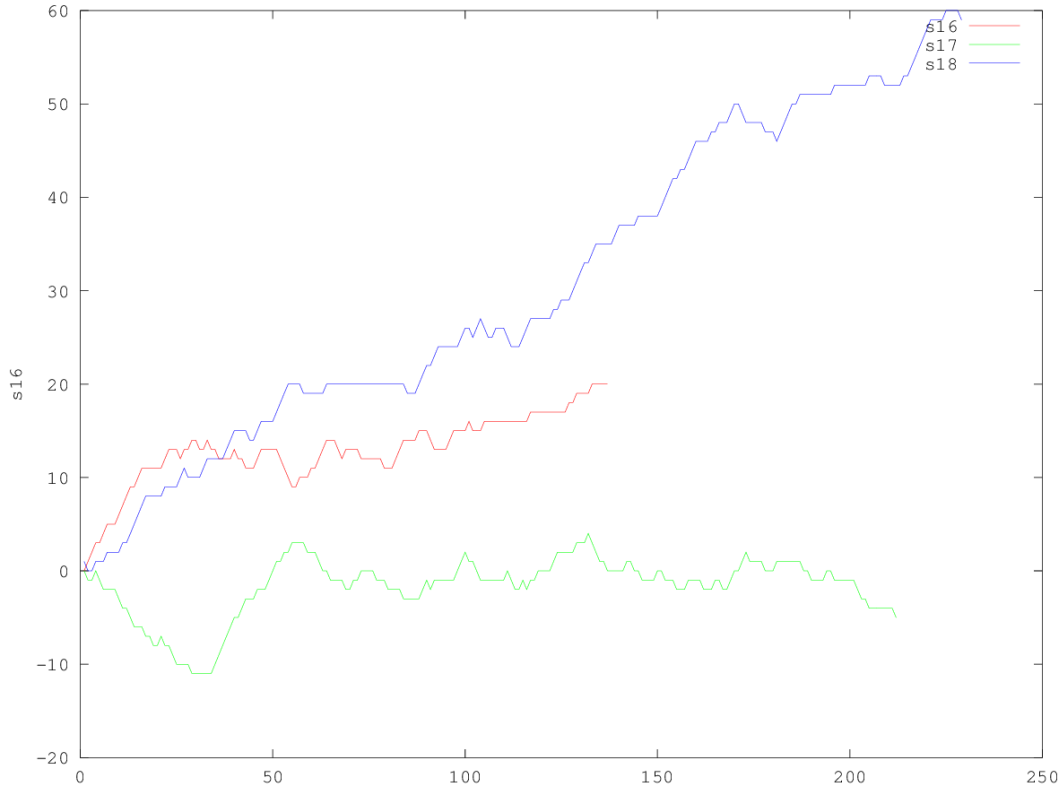
- **Group 12**: ARS=-1.819



- **Group 9**: ARS= -1.849



- **Group 6**: ARS= -2.432



One thing we notice is that in the groups that performed badly, certain members gain a very high authority score, despite the number of conversations being comparable to the good groups. This indicates that they issued a lot of K1 type statements. The other members in such groups have significantly lesser participation (they make much lesser number of moves). Moreover, the high authority score despite the same number of moves as good groups suggests a dearth of K2 moves. It is also possible that most of the K1s that occurred are proposals rather than clarifications since clarifications usually occur in response to a doubt. However, this cannot be said for sure, since in integration based consensus building, a K1 following a proposal by another person, would function as a clarification/elaboration. This is why we cannot directly make a statement about consensus or confusion levels by looking only at the authority graphs.

10 Full Automation

As a possibility of complete automation, we experimented with a cutting edge automatic sequencing procedure [n-12]. It implements the first two steps of the pipeline proposed in this paper, namely: a) Negotiation Labeling b) Sequencing For the negotiation labeling, the same classier mentioned before was used, with a kappa of 0.7022. As for the sequencing, the precision and recall for the automated sequencing were 0.746 and 0.698 respectively. The f-score obtained was 0.7208. As our dataset is small by conventional standards, a 5 fold cross-validation was carried out. However, despite the seemingly good f-score, using this automated sequencing along with the automatically generated tags for generating active learning sequences degraded performance (kappa of 0.36 with respect to the human annotations), mostly because the automated sequencing uses cosine similarity and textual features to decide whether to assign a new conversational move to an existing sequence or

create a new sequence. This approach is prone to favor creation of new sequences, since there are no special textual features (unigrams/bigrams) associated with the start of a new sequence. So any features found in new sequence starters are likely to be present in non-sequence starters as well. Starting a new sequence is usually due to a combination of a change in topic, speaker or a large time delay since the previous sequence etc. rather than the occurrence of specific unigrams or bigrams.

11 Future Work

In our work so far, we have identified an automatable pipeline for chat analysis that exploits information about changes in authority levels of speakers within sequences to identify stylistic patterns in conversation that have been shown to be associated with learning in past research, and have mapped them two quantitative metrics of consensus and confusion. In this pilot study, we have explored the possibility of automation of the pipeline proposed, using a semi-supervised method (that relies on manual annotation). The results are encouraging and indicate that in light of cutting edge work in automated chat segmentation and sequencing [15] there is a real possibility of a fully automated chat analysis procedure in the future. Our future work would be in this direction. More specifically, we shall explore expanded feature sets to improve automated sequencing performance. Some work in this area [25][26] does indicate that this could be beneficial. In addition to this, our future efforts would also involve the adaptation of this trajectory based analysis of conversation, to a real time system that can be used to influence tutor interventions in a mediated collaborative learning setting. Some observations regarding this have already been made in the results section. Formalizing and operationalizing this is an exciting future prospect.

References

[1] France Henri, Computer Conferencing and Content Analysis. Series F : Computer and Systems Sciences, Springer Berlin Heidelberg, 1992

[2] B.D Wever, T. Schellens, M. Valcke, H. Van Keer : Content Analysis Schemes to Analyze Transcripts of Online Asynchronous Discussion Groups, Computers & Education, Volume 46, Issue 1, January 2006, pages 6-28

[3] K A Meyer, Evaluating Online discussions : Four Different Frames of Analysis, Journal of Asynchronous Learning Networks, 2004

[4] P J Fahy, G Crawford, M Ally : Patterns of Interaction in Computer Conference Transcript, The International Review of Research in Open and Distance learning, 2001

[5] Armin Weinberger, Frank Fischer, A framework to analyze argumentative knowledge construction in computer-supported collaborative learning, Computers & Education, Volume 46, Issue 1, January 2006, Pages 71-95

[6] Leitao, The potential of knowledge building, Human Development, Vol. 43, No. 6, 2000

[7] Stephanie D. Teasley, Talking about Reasoning : How Important is the Peer in Peer Collaboration, Discourse Tools and Reasoning : Essays on Situated Recognition, 1997

[8] Gunawardena, Lowe & Anderson, Analysis of Global Online Debate And the Development of an Interactive Analysis Model for Examining Social Construction of

Knowledge in Computer Conferencing, Journal of Technical Writing and Communication, Baywood Publishing Company, 1997

[9] Martin and D. Rose, Working with Discourse : Meaning Beyond the Clause, 2003

[10] Amy Soller & Alan Lesgold, Analyzing Peer Dialogue from an Active Learning Perspective Proceedings of the AI-ED 99 Workshop: Analysing Educational Dialogue Interaction: Towards Models that Support Learning, LeMans, France, 63-71.

[11] Christopher M. Mitchell, Kristy Elizabeth Boyer, James C. Lester : From Strangers to Partners : Examining Convergence within a Longitudinal Study of Task Oriented Dialogue, SIGDIAL 2012

[12] PW Foltz, MJ Martin, A Abdelali, MB Rosenstein, RJ Oberbreckling : Automated team discourse modeling: Test of performance and generalization, Proceedings of the 28th Annual Cognitive Science Conference,2006

[13] Amy Soller, Janyce Wiebe, Alan Lesgold : A Machine Learning Approach to Assessing Knowledge Sharing During Collaborative Learning Activities, Proceedings of Computer-Support for Collaborative Learning 2002, Boulder, CO, 128-137.

[14] Amy Soller, Alan Lesgold : A Computational Approach to Analyzing Online Knowledge Sharing Interaction, Proceedings of Artificial Intelligence in Education 2003, Sydney, Australia, 253-260

[15] Elijah Mayfield, David Adamson, Carolyn Penstein Rose : Hierarchical Conversation Structure Prediction in Multi-Party Chat, SIGDIAL 2012

[16] M.M Chiu. Group Problem Solving Processes : Social Interactions and Individual Actions, In theory of Social Behavious, 2000

[17] D. Suthers, N. Dwyer, R. Medina and R. Vitrapu : A Framework for Eclectic Analysis of Collaborative Interaction, In International Conference on Computer Supported Collaborative Learning, 2007

[18] Elijah Mayfield, David Adamson, Alexander I. Rudnicky, Carolyn Penstein Rose :Computational Representational of Discourse Practices Across Populations in Task-Based Dialogue, ICIC, Bangalore, 2012

[19] Iris Howley, Elijah Mayfield, Carolyn Penstein Rose : Missing Something? Authority in Collaborative Learning, Computer Supported Collaborative Learning, 2011

[20] Hinze, U., Bischoff, M. & Blakowski, G. (2002). Jigsaw Method in the Context of CSCL. In P. Barker & S. Rebelsky (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2002 (pp. 789-794). Chesapeake, VA: AACE.

[21] Beatriz Barros & M. Felisa Verdejo : Analysing student Interaction processes in order to improve collaboration – The DEGREE approach , International Journal of Artificial Intelligence in Education 11 (3), 221-241

[22] Elijah Mayfield and Carolyn Penstein Rosé : An Interactive Tool for Supporting Error Analysis for Text Mining , In the Demonstration Session of NAACL 2010.

[23] Deanna Kuhn : Skills of Argument, Cambridge University Press,1991

[24] *Spiro, R. J., Feltovich, P. J., Jacobsen, M. J., Coulson, R. L. (1991):* Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains. Educause Review, 5, pp. 24–33.

[25] *Joonsuk Park, Claire Claudie:* Improving Implicit Discourse Relation Recognition through Feature Set Optimization, SIGDIAL 2012

[26] *Brad Goodman, Frank Linton, Guido Zarrella, Robert Gaimari:* Using Machine Learning to Predict Trouble During Collaborative Learning