Using Transactivity in Conversation for Summarization of Educational Dialogue

Mahesh Joshi[1] *and Carolyn Penstein Rosé*[1,2]

[1] Language Technologies Institute, [2] Human Computer Interaction Institute
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
{maheshj,cprose}@cs.cmu.edu

Abstract

We present our ongoing work towards using the concept of transactivity [1] for automatically assessing learning of students working together in a collaborative setting. Transactive segments of student dialogue are proposed as useful components of conversation summaries generated for instructors. Experimental evaluation of this hypothesis shows promising results. Further, initial results are presented for automatic identification of transactive contributions in student dialogue.

Index Terms: transactivity, conversation summarization, educational dialogue

1. Introduction

The task of a group-learning facilitator is to monitor a large number of on-going collaborative learning discussions and to intervene when necessary to keep the conversation moving in a productive direction [2]. In order to do this job successfully, the facilitator must be strategic in selecting which groups to be involved in [3]. Ideally, the facilitator should be able to quickly form an impression of how effectively students are interacting with one another in order to identify which groups need help the most. We refer to the task of generating conversation summaries to aid group-learning facilitators in identifying groups that need help as the Group Learning Facilitator (GLF) task.

It has been found that effective learning in collaborative groups is linked to the process by which learners work on the learning task together [4], how they construct arguments and argumentation sequences [5], [6], and how they build on the contributions of their learning partners [1], [7], otherwise known as transactivity. We leverage this theoretically motivated idea of transactivity in discourse, and integrate it with a design methodology for conversation summarization.

The remainder of the paper describes the motivation and the three stages of the proposed approach to conversation summarization. The baseline approach we use for comparison is a purely empirical, bottom-up approach, which we demonstrate to be less valuable for supporting group learning facilitators. We validate our approach by requiring group-learning facilitators to rank students based on how much they learned, in order to identify those students who could have used more support. An effective summary to support this task would allow a group-learning facilitator to quickly and accurately rank students according to which ones need the facilitator's help most. We measure the accuracy of their ranking by comparing it with a gold standard ranking determined by pre- to post-test gains. Finally we present our work in progress towards automatic classification of dialogue contributions as transactive or non-transactive, which is an intermediate step for construction of summaries for the GLF task. We include an error analysis and discussion of current directions.

2. Motivation

Participants in a collaborative learning setting are said to have a transactive discussion when they elaborate, build upon, question, or argue against the ideas presented by their partners in the process of working towards a common understanding of the task and reaching a shared solution. This process of understanding the partners' ideas, comparing them to one's own understanding, arguing and forming a common ground upon which a solution can be built collaboratively has been shown as important for collaborative learning [7], [8].

Transactivity is well studied in the domain of educational psychology and Computer Supported Collaborative Learning (CSCL). Berkowitz and Gibbs [1] provide an extensive categorization of transactive contributions with several concrete examples in each category. They refer to a transactive contribution in a discourse as a "transact". Some examples of transacts from their coding scheme are listed in Table 1. Unlike the informational relations within Rhetorical Structure Theory [9] that are primarily meant to connect the pieces of one speaker's argument, transacts are specifically formulated to represent the relationship between competing positions of different speakers. Although transacts are defined at the same level of discourse granularity as that of "dialogue acts" [10] that assign a semantic category to dialogue contributions (such as a question or answer), their semantic granularity differs. For example, a dialogue act such as a "question" could be any of the following transacts – a "feedback request", a "justification request" or even an "extension" phrased as a question, the difference among these transacts being the extent of understanding or comprehension that they signify. These distinctions are often important for identifying which episodes within a collaborative learning interaction were responsible for its success.

Table 1. *Examples of transacts and their condensed definitions.*

Transact	Definition
Feedback Request	Do you understand or agree with my position?
Dyad Paraphrase	Here's a paraphrase of a shared position.
Competitive Juxtaposition	I will make a concession to your point, but also reaffirm part of my position.
Extension	Here's a further thought or an elaboration offered in the spirit of your position.

Coding schemes such as the Berkowitz and Gibbs scheme can be used to identify key segments of discourse that have predictive value for an outcome of choice, in our case learning gains. Once it is determined which aspects of discussions have predictive value, the model can be used to select portions of an extended interaction that are strategic to include in a summary,

which can then be used to support a human in the task of making a prediction about the outcome of choice. Here is an example of a transactive exchange of the *Competitive Juxtaposition* variety:

A: well …u do know increasing tmax and pmax mean more Qin

B: yeah - but more quality - which means you get more work out of the turbine

Rather than present to facilitators a condensed version of the interaction meant to convey purely the information content of a conversation, the excerpts selected and included in the summary are meant to give a sense of the nature of the conversation, since the nature of that interaction, such as how transactive it was, is an important piece of evidence for predicting the effectiveness of the interaction for supporting learning.

3. Methodology and Results

The methodology is presented in three stages, starting with the design of a conversation summary, followed by an empirical validation of the design, and finally reporting on the initial results towards automatic summary generation.

3.1. Stage 1: Design of a Conversation Summary

The goal of this stage is to determine which characteristics of the conversations need to be conveyed in the summaries. The predictive value of a characteristic is evaluated by checking for a statistically significant correlation between the occurrence of that characteristic in a conversation and the value of the outcome of interest (in our case – learning gains). For example, assume it is possible to determine the number of times students asked their partner a question. If there is a significant correlation between this number and how much the student learned, then one would hypothesize that conveying how many questions a student asked, along with some examples of key questions, might be valuable to include in the summary.

We hypothesize that transactivity in collaborative discourse would be a better predictor of learning as compared to a purely bottom-up approach based on automatic corpus analysis. Thus transactivity is proposed as an intermediary that connects evidence from discourse to learning gains for students.

In order to evaluate this hypothesis, we have analyzed instant messaging chat logs of students working in pairs on solving a Thermodynamics design problem. Specifically, the task that the students worked on is the design of thermodynamic cycles in engines, to optimize the cycle performance. We have a total of 24 chat logs from 48 students. Each chat log consists of the typed contributions from two students who formed the pair. There is a wide variation in the number of total contributions in a dialogue, ranging from as low as 16 to as high as 139.

The bottom-up automatic analysis of chat contributions poses the task as a regression problem of predicting the learning gains for each student based on several features that are extracted from the chat logs after segmenting each dialogue into on-topic and off-topic chunks. The features extracted for each chunk are listed in Table 2. For all our automatic evaluations, we have used the leave-one-dialogue-out testing strategy to obtain the most valid evaluation metrics. After obtaining the learning gain predictions, the students are ranked according to the predictions and these ranks are compared to the true ranks obtained using the actual pre- to post-test learning gains. The ranking correlation using different sets of features is shown in Table 3. Best results for regression using Support Vector Machines (SVMs) [11] with a linear kernel are reported, although we have tried several other regression techniques that did not perform as well (except when including stopwords as features, row 5 of Table 3, where SVM with a Gaussian kernel gave a correlation of 0.071). This demonstrates the low utility of baseline features for our task.

Now we turn our attention to evaluate the utility of transactivity based features. For this evaluation, the chat corpus was annotated using the coding manual developed by Berkowitz and Gibbs [1], to categorize each contribution in the 24 dialogues into one of the low level transacts. A special label "NONE" was assigned to non-transactive contributions. We had

Table 2. *Features used for bottom-up regression approach.*

Group	Feature
Lexical / Pseudo-syntactic (the last 3 grouped as "Misc" in result tables)	Unigrams
	Bigrams
	Part-of-Speech bigrams
	Punctuation
	Normalized contribution length
	Contains at least one non-function word?
Meta	Is self's contribution on-topic?
	Is self the initiator in the chunk?
	Is self the major contributor in the chunk?
	Proportion of numbers in self contribution
	Cosine similarity between self and partner's portion of the chunk
	Relevance [12] of self to partner
	Informativity [12] of self w.r.t. partner
	Self-performance on pre-test High/Low?
	Partner performance on pre-test High/Low?

Table 3. *Ranking correlation of SVM regression with true ranking. Each row includes features from all the previous rows, except last row where stopwords are removed.*

Features	SVM (Linear Kernel)
Uni	0.177
+Bi	0.116
+Pos-Bi	0.103
+Misc	0.155
+Stopwords	0.039
+Meta -Stopwords	0.157

a total of 1,580 contributions and the annotation led us to a total of 180 transacts in different categories. On a subset consisting of 139 contributions, we calculated kappa agreement for human evaluation on a binary transact vs. non-transact categorization task, which came to 0.71.

In addition to evaluating the predictive power of individual transacts, we also evaluated five high level categories, which represent clusters of similar transacts, namely ego-oriented, alter-oriented, dyad-oriented, competitive and non-competitive transacts and counted the number of transacts for each student in those categories. We computed regression models using the learning gains of students as the dependent variable and the percentage of different types of transacts in a student's contributions as predictors. Table 4 summarizes the best results.

Correlation coefficient values in Table 4 can be directly compared to the ones presented earlier using the bottom-up regression approach since the dependent variable is the same in

both cases. Notice that results using transactivity based predictors are a substantial improvement, although the value is still low. Two conclusions can be drawn from this analysis: 1) the design based on transactivity analysis of the conversations works better than a regression approach based purely on low-level lexical and pseudo-syntactic features. 2) Specifically among the transact categories, we find that transacts that relate to reasoning of both the students in the group (dyadic transacts) are significantly correlated with learning. Dyadic transacts are evidence that a student is comprehending his/her partner's reasoning as well as incorporating one's own reasoning into partner's position, either competitively or non-competitively.

Table 4. *Regression analysis for learning gains of students.*

Transact	Correlation
Dyad Paraphrase	0.247 ($p < 0.10$)
Non-competitive Dyad	0.251 ($p < 0.10$)
Dyad	0.297 ($p < 0.05$)

3.2. Stage 2: Empirical Validation of Design

The result of the investigations at the design stage indicated the potential value of using transactivity as a basis for summarizing learning interactions for the GLF task. The purpose of the validation stage is to determine whether using transactivity as a basis for selecting portions of discussion transcripts for a summary is valuable to facilitators.

In order to measure the potential value of using transactivity as a basis for summarization, we set up an experiment to compare how well humans are able to rank students in terms of expected learning based on the raw chat transcript from their group discussion with how well they are able to rank students when the transactive regions of the dialogue are highlighted. Accuracy was assessed by computing a correlation between the rankings assigned by the human judges with gold standard rankings based on actual learning gains. Note that this correlation is equivalent to those presented in Table 4 above since in both cases the gold standard rankings are based on pre- to post-test gains. If the ranking is more accurate using the highlighted transcripts, then it can be concluded that transactive segments in discourse are valuable components for inclusion in a conversation summary. We set up two piles of a subset of 32 transcripts (out of the total 48 where each transcript is repeated for both the students in the pair) so that within each pile there was one transcript per student, and that student's contributions to the dialogue were labeled with that student's identification tag in boldface type. The other contributions were labeled with the partner's contribution tag in regular type. In one pile of 32 transcripts, all of the transactive contributions of the student indicated in boldface type were highlighted. In the other set, no such highlighting was provided.

In order to make a fair comparison we took precautions against ordering effects, time-on-task confounds if one method of ranking turned out to be faster, or biases from individual humans. In order to achieve this, each of two human judges received a set of 32 transcripts, half of which had the trans-activity highlighting and the other half of which did not. The transcripts were divided into 4 piles of 8 transcripts each, alternating between a pile with highlighting and a pile without. The two judges were reversed with respect to which type of pile they started with.

Then the human judges did the ranking task as follows. They spent 15 minutes on each of the 4 piles in the order in which they were given the piles. They then spent 15 minutes combining the 1st and 3rd piles, which for one judge were the two highlighted piles, and for the other judge were the two non-highlighted piles. Finally they spent 15 minutes combining the 2nd and 4th piles. Thus, for each judge we ended up with two piles, each with a ranking of half of the students. For each judge we then computed a correlation between their ranking and the gold standard ranking for each of their two piles, one of which being the pile with highlighting and the other of which being the pile without high-lighting. We then averaged the correlations with highlighting across the two judges to obtain a correlation coefficient of 0.299. Similarly, averaging the correlations without highlighting across the two coders obtained a correlation coefficient of 0.018. Thus, the human judges achieved roughly a 0 correlation with the gold standard ranking when they did not have the aid of the transactivity highlighting. They achieved a correlation coefficient of 0.299 when they had the highlighting, which is a substantial improvement. Note that the comparison between human performance with and without transactivity highlighting is analogous to the comparison between the fully automatic models using either low level linguistic features or including predictors derived using a transactivity analysis, which was presented above.

3.3. Stage 3: Automatic Transact Classification

As shown earlier, highlighting transactive contributions of a discourse is useful for ranking students. In order for a summary to include such transactive contributions, the summarization system should successfully identify transactive contributions from non-transactive ones. Further, in the ideal case the transactive contributions should be categorized into one of the low-level or high-level categories such as the dyadic transacts. To this end, we present our ongoing work towards automatic identification of transacts.

The task in this case is a text classification problem: given a student contribution, automatically decide whether it is a transactive contribution or not. We have done some initial experiments using two off-the-shelf classifiers: SVMs [11] and Collins' sequential Voted Perceptron Learner [13]. We performed leave-one-dialogue-out evaluation, using the MinorThird toolkit [14] for implementations of the above algorithms. The best feature set we identified for these experiments was unigrams, bigrams, Part-of-Speech bigrams and the "Misc" group of features, with function words included. We evaluated the importance of function words for this task in two ways: 1) the straightforward way by comparing performance when excluding vs. including the function words (rows 4 and 5 in Table 5 respectively) and 2) by treating a semi-automatically generated list of domain words as a list of function words and therefore eliminating them, while keeping the function words as features (row 6 in Table 5). It can be seen that excluding function words is at least as much or more detrimental than excluding the domain words, especially for our SVM models. Note that although we are using features that were unsuccessful for making a general prediction about learning effectiveness from a whole dialogue, it is still reasonable to attempt to use features such as these to assign the much more specific transactivity based categories to specific regions of the conversation.

We report our results in terms of the Cohen's Kappa value and F1 measure (which treats precision and recall equally) on

identifying transacts. Table 5 shows the results for different feature sets described above.

Table 5. *Results on binary transact classification.*

Feature Set	SVM		Collins' Perceptron	
	Kappa	F1	Kappa	F1
Uni	0.282	0.336	0.291	0.351
+Bi	0.317	0.375	0.292	0.353
+Pos-Bi	0.354	0.424	0.416	0.480
+Misc	0.374	0.442	0.407	0.470
+Stopwords	0.393	0.460	**0.453**	**0.506**
-Domain words	0.399	0.466	0.408	0.470

3.3.1. Error Analysis

For our best result, a major part of the error was due to low recall of 0.444 (the precision was 0.588). There were many false negatives. Analysis of our best model (the Collin's Perceptron)

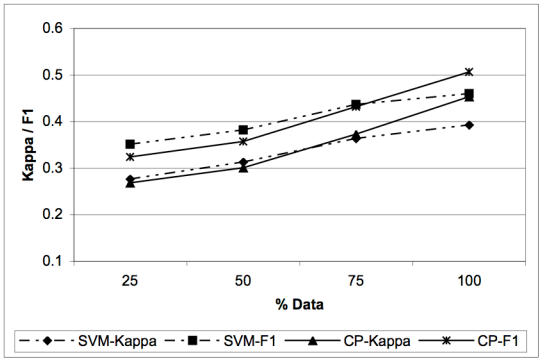

Figure 1. Learning curves for SVM and Collins' Perceptron.

showed a highly negative weight for period (the punctuation mark), which might be a peculiarity for this dataset. Highly negative weights were learned for features that take into consideration whether previous contributions are transacts, which is actually counter intuitive as many times transacts follow each other in our data. Nevertheless, such instances got misclassified due to these high negative weights. This might have been due to insufficient training data to learn sequential patterns of transacts. We evaluated this hypothesis by plotting a learning curve based on using 25%, 50%, and 75% of the data for our leave-one-dialogue-out evaluation. The curve is shown in Figure 1 and shows a more-or-less steady increase in performance with data even from 75% to 100%, indicating that it might be boosted further with more data.

4. Conclusions and Current Directions

We have presented a 3-stage design approach for conversation summarization. In the first stage of summary design, we propose the use of transactivity of discourse as a means for identifying segments of discourse that help in predicting effective learning happening in group discussions. We validate this design by evaluating how knowledge of transactive portions of a discourse helps humans to better distinguish students who are learning effectively from those who are not. In the third stage, we report our work in progress to automatically identify transactive contributions in a discourse, which shows encouraging results even with the limited amount of data that we are using.

Current directions for this work include use of deeper syntactic features in combination with contextual knowledge to improve our performance. The utility of syntactic features is evidenced by a considerable improvement using simple pseudo-syntactic features such as Part-of-Speech bigrams. Identifying the right context that helps in classifying a contribution is important since transactivity is by definition building upon partner's reasoning, which should be extracted from previous context. We also plan to evaluate information extraction approaches to identify domain relevant causal phrases from discourse, which might be useful features for predicting transactive contributions.

Acknowledgements: This work was supported by Office of Naval Research, Cognitive and Neural Sciences Division, Grants N00014-05-1-0043 and N00014-04-1-0107.

5. References

[1] Berkowitz, M., & Gibbs, J. Measuring the developmental features of moral discussion. Merrill-Palmer Quarterly, 29, 399-410, 1983.

[2] Hmelo-Silver, C. E. & Barrows, H. S. Goals and Strategies of a Problem-based Learning Facilitator. Interdisciplinary Journal of Problem Based Learning, 1(1), 21-39, 2006.

[3] Soller, A., and Lesgold, A. A Computational Approach to Analyzing Online Knowledge Sharing Interaction. Proceedings of Artificial Intelligence in Education, 2003.

[4] Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. Fostering collaborative knowledge construction with visualization tools. Learning and Instruction, 12, 213-232, 2002.

[5] Leitão, S. The potential of argument in knowledge building. Human Development, 43, 332-360, 2000.

[6] Voss, J.F. & Van Dyke, J.A. Argumentation in Psychology. Discourse Processes, 32(2&3), 89-111, 2001.

[7] Teasley, S. D. Talking about reasoning: How important is the peer in peer collaboration? In L. B. Resnick, R. Säljö, C. Pontecorvo & B. Burge (Eds.), Discourse, tools and reasoning: Essays on situated cognition, 361-384, 1997.

[8] Azmitia, M., & Montgomery, R. Friendship, transactive dialogues, and the development of scientific reasoning. Social Development, 2(3), 202-221, 1993.

[9] Mann, W.C., & Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281.

[10] Stolcke, A., Ries, K., Coccaro, N., Shriberg, J., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C. & Meteer, M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3), 339–373, 2000.

[11] Vapnik, V. The Nature of Statistical Learning Theory, 2000.

[12] Olney, A. & Cai, Z. An Orthonormal Basis for Topic Segmentation in Tutorial Dialogue. In Proceedings of HLT-EMNLP 2005, 971–978, 2005.

[13] Collins, M. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In Proceedings of EMNLP, 2002.

[14] Cohen, W. Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net, 2004.