Experiments to Classify and Measure Conversational Interactivity in COMPS Problem-Solving Dialogues

Michael Glass
Melissa Desjarlais
Valparaiso U

Jung Hee Kim
Kelvin Bryant
NC A&T State U

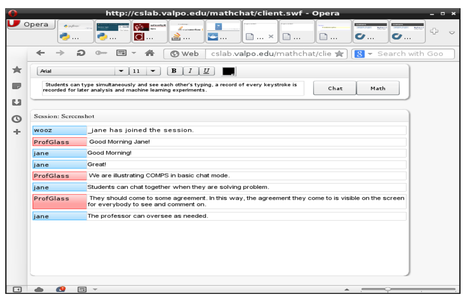
Abstract

This paper reports on experiments in identifying whether students are responding to each other and measuring the general level of conversational interactivity in COMPS problem-solving dialogues. COMPS is a web-delivered computer-mediated problem solving chat environment for student collaborative exploratory learning.

- We focus on the Initiate (I) and Respond (R) construct from Conversation Analysis Exchange Structure theory. More interactive (and more transactive) conversations should exhibit a higher fraction of R turns.
- We attempted to train models to:
  a) classify individual turns as I or R.
  b) measure the general level of conversational interactivity by predicting the percentage of R turns.

Where the Dialogues Came From

- This work has focused on transcripts from a 2nd year Java programming class.
- Students work in 3 or 4 person groups solving problems in understanding Swing GUI principles.
- Students converse until a shared understanding of the answers is achieved.
- Then they see the correct answers, and converse until shared understanding is again achieved.
- Logs contain student dialogue text plus time stamps for every keystroke.
- All features for classifiers are mechanically extracted from text or timing data.



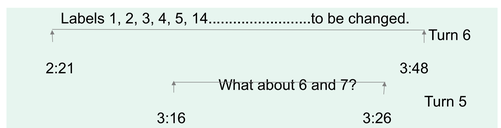
Sample Marked-Up Dialog

Dialogue extracted annotated for Initiation and Response

Turn	Stu	Text	Annotation	Start time	End time
1	C	hey people		00:00	00:02
2	C	okay question one??	I	00:43	00:47
3	B	I'm reading it	R	01:08	01:15
4	C	do either of you know what the question is even asking? i don't	I	01:44	02:10
5	B	what about 6 and 7?	R	03:16	03:26
6	A	"Labels 1, 2, 3, 4, 5, and 14 can be instantiated anonymously. Because these do not have to be changed."	R	02:21	03:48
7	B	that makes sense	R	04:12	04:17
8	A	6 and 7 can not be instantiated anonymously because these values have to change.	R	03:59	04:18
9	C	okay. Im lost where are you guys gettting this from	I	04:19	04:43
10	C	the back ground information?		04:46	04:53
11	A	It's on the second page.	R	04:52	04:56
12	B	the top discription	R	04:58	05:02
13	C	ohhh mow i see thanks	R	05:08	05:16

Timing difference features record (illustrated by A=turn 6 and B=turn 5)
Astart-Bend: -65sec
Astart-Bstart: -55sec
Aend-Bend: +22sec
Aend-Bstart: +32sec

Each turn by participant A produced two records: A vs B's most recent turn and A vs C's most recent.



Sequence inversion caused by simultaneous typing.

Turns 7 and 8 exhibit similar inversion.

Experiment: Classify I/R

Session Statistics

Sessions	17
Dialogue Turns	1827
Turns per Session	107
Median Duration (min)	52
Shortest, Longest (min)	26-67

Interactivity: R / (I+R)

Turns marked I or R	1790
Mean of all turns	0.65
Mode of 17 sessions	0.64
Minimum session	0.49
Maximum session	0.72

From the texts extract features that can be used for machine learning.

- The presence of discourse marker words such as *so*, *therefore*, *now*, often accompany reasoning statements or topic shifts.
- The presence of words in the Java Swing problem domain such as *text field* and *mouse listener*.
- Whether or not two people are typing at the same time.
- Deictic references (e.g. pronouns and names) to parts of the problem: *label 1*.
- Emoticons ;-)
- Question marks.
- Pronouns such as *you* and *we* that indicate more than one person are involved in this exchange.
- Timing differences: e.g., how long after person A stopped did person B start? Short times are associated with replies.
- Length of turns: one to three word turns often answer questions or are simple "ok" acknowledgments.

Prevalence of Some Text Features (% of Turns)

Discourse markers	10%
Problem domain words	20%
Overlapped typed turns	47%
Task-related deixis	30%
Emoticons	1%
Question marks	14%
Pronouns	16%

Results

Classifier for individual turn I or R.

- Trained J48 decision trees (because many of the features are binary)
- With and without timing differences
- Several editions of hand-annotated transcripts

Result:

- Cross-validated Kappa between machine and human typically 0.26
- Most predictive feature: inter-turn time Astart - Bend

Predict interactivity R/(I+R) of each session.

- Counted binary features to create numerical features
- Used multiple linear regression
- Single time difference: time since most recent turn by any other participant

Result:

- Cross-validated R correlation between fitted line and data is close to 0

Discussion

Why are we not predicting I/R very successfully?

It is possible our I/R markup is flawed.

- Our I/R is not rooted in any one theory or existing markup manual.
- Results on 3 sessions that were annotated in multiple passes (according to our own social mode dimension) were noticeably better than overall results on all 17, of which most were annotated by one rater only.

Timing anomalies abound. 47% of typing occurs while other students are typing.

- We did not appreciate the level of full duplex communication that our students engage in until we started extracting features for this experiment.
- Most chat systems do not permit students to see each other's words and type simultaneously as COMPS does.
- In spoken dialogue this kind of full duplex communication is also not possible.
- We don't know how to analyze it yet. Neither does anybody else that we know of.

Conclusions

Conclusion 1: Not yet.

Conclusion 2: Redoing this and similar experiments in a different way might yet be fruitful.

Conclusion 3: Full duplex student interactions could be very interesting.

Background: I (initiate) and R (respond)

Why I and R?
Conversation analysis (a discipline of Linguistics) recognizes *exchange structure*, segments of conversation that start with one person initiating and continue with participants responding and possibly following up.

In educational dialogue, I and R are useful for recognizing phenomena such as whether students are responding to each other's reasoning.

Alice:	Your line was busy.	(Initiate)
Bob:	Sorry, Carol called from school.	(Respond)
Alice:	OK	(Followup)
Bob:	Are you concerned about her also?	(Initiate)
Alice:	I was concerned whether we are prepared for the fall semester.	(Initiate)
Bob:	Ah. I've done what you asked.	(Respond)
	And I have a cute new homework.	(Initiate)

1	C	i think its only 1-5
2	A	why only 1-5
3	C	and 14
4	B	yeah i dont understand why it would just be 1-5, 14
5	C	because 6 7 are being updated at different times
6	B	yeah right, i didn't even look at the top of the page
7	A	youre right
8	B	Yes okay, 1-5 and 14
9	B	For #2 would be mouse listener...
10	A	i think number 3 is 8,9,10
11	C	you can pick more than one for number 2

We made our own rules for I/R:

- Multiparty conversations cannot be easily segmented.
- Other useful discourse analyses look at whether (and how) an utterance engages with previous utterances:
  - Transactivity -- the social mode of knowledge construction
  - Centering -- identifying which NPs are candidates for pronominalization

We called a turn R if there was a common reference or idea and it "responded" to the earlier reference or idea.

Fun exercise: classify the turns in the dialogue to the left.

Acknowledgments

Thank you to our hard-working students at North Carolina A&T and Valparaiso Universities, particularly Micayla Goodrum and Thomas Martin.

This work was supported by the National Science Foundation under award 0633953 to North Carolina A&T State University and 0634049 to Valparaiso University.