Making Sense of Large-Group Discussions using Rhetorically Structured Text

Ana Cristina Bicharra Garcia
Computer Science Department
Universidade Federal Fluminense
Niterói, Brazil
bicharra@ic.uff.br

Mark Klein
Sloan School, MIT
Boston, USA
University of Zurich
Zurich, Switzerland
m_klein@mit.edu

***Abstract*— Recent advances in social media technology have made it possible to involve large groups in online deliberations, using such tools as forums and argument maps. As discussions develop, however, making sense of their content can become a big challenge for newcomers, thus impeding their potential participation. We posited that rhetorically organized narratives can foster superior comprehensibility, and conducted an experimental evaluation that supports this claim. Human subjects were asked to answer a questionnaire about a discussion presented in one of three formats: forum, argument map, and rhetorically structured text. The rhetorical structures produced superior question-answering performance for complex questions. In this paper, we discuss these results, as well as their implications for the design of large group interaction tools.**

Keywords-Rhetorical structure theory; RST; online group deliberation; web forums; argument maps; crowd-computing; social computing

I. INTRODUCTION

Forums are virtual places for hosting online discussions among people on subjects of mutual interest. They have been an important source of knowledge in many fields, ranging from computer software development to education [1]. Typical forums have a *chronological* structure. Each new contribution (post) is appended at the end of the list of previous contributions, labeled by a time stamp and the name of its author. As a discussion develops, however, it becomes increasingly difficult for newcomers to understand the intertwined contributions from other participants, and this problem gets amplified as the group grows. Threaded discussions provide an additional layer of organization, based on capturing post reply structures [2], but this is of limited value for increasing comprehensibility because there is no clear relationship between reply structures and the semantics of a discussion.

Argument maps [3]-[7] provide an alternative, logic-based, structure where users organize their contributions using a pre-defined taxonomy of post types (e.g., issues, ideas, pros and cons). Such maps reveal the intentions of the posts, but at the cost of removing information about the chronological order of contributions, potentially impairing understandability [8].

In this paper, we explore whether a *narrative* organization for discussion content, one based on the principles of rhetorical structure theory (RST), can transcend the comprehensibility limitations of (chronologically-structured) forums and (logically-structured) argument maps. To test this idea, we conducted an experimental comparison with three groups of 16 people in each. Participants in the groups were demographically balanced by gender, age and background. Each group interacted with the same discussion content in one of these three formats:

- discussion forum
- argument map
- rhetorically-structured text

We measured how quickly and well the participants in each group could answer a range of questions about the content. We found that rhetorically structured text significantly improved the participant's abilities to answer complex questions relative to web forum and argument map structures ($p<0.05$), but did not have a statistically significant impact on answering simple questions.

Section 2 presents related work, followed by background explanation concerning RSTs presented in Section 3. Section 4 describes the experiments and finally Section 5 presents the conclusions including final remarks and future work.

II. RELATED WORK

The goal of crowd-scale deliberation is to allow communities to identify and evaluate possible solutions for a problem of shared concern [9][10]. A wide range of social computing technologies have emerged to address this challenge in the past few decades, including email, chat, wikis, web forums, open innovation systems, group decision support systems, as well as debate and argumentation systems.

How well do existing social computing technologies fare in terms of realizing these potentially powerful effects in the context of crowd-scale deliberation? There are several key types of applicable technology, each with their own strengths and weaknesses, including time-centric systems, question-centric systems, topic-centric systems, debate-centric systems, and argument-centric systems. We will review each type in the sections below.

 ISBN: 978-1-61208-443-5

A. Time-Centric Systems

Time-centric systems include tools - such as, email, chat rooms, blogs, micro-blogs like twitter, and web forums – in which content is organized based on when a post was contributed. Currently, time-centric systems are by far the dominant technology used for online deliberation. These systems enable large communities to weigh in on topics of interest, but face serious shortcomings that can deeply undercut the value of the deliberation engagements [11], such as:

- Low signal-to-noise ratios,
- Insular ideation,
- Balkanization,
- Non-comprehensive coverage,
- Dysfunctional argumentation and
- Opaque Processes

Because of all these issues, the content generated by time-centric deliberation tools is typically very sub-optimal from both a depth and breadth perspective.

B. Question-Centric Systems

Question-centric systems [12] are organized around questions: one or more questions are posted and the community is asked to contribute, rate, and comment on proposed solutions for these questions. These systems can be divided into two subtypes based on whether the questions are "close-ended" (there are only one or few correct answers, and the answers are relatively easy to verify), and "open-ended" (the system is soliciting ideas for large complex problems which have many possible solutions and where identifying the best answers is not straightforward). Close-ended question-centric sites such as stackoverflow.com, a programming Q&A site, have been remarkably successful [13], but are applicable only to a small subset of the entire scope of potentially important deliberation problems. Open-ended systems - such as group decision support systems as well as such open innovation platforms as IdeaScale and MindJet - can elicit huge levels of activity, and organize content better than time-centric tools. Like time-centric systems, however, they are prone to high levels of redundancy, wherein many of the ideas represent minor variations of each other. Also like time-centric systems, they tend to elicit many simple single-author ideas rather than a smaller number of collaborative efforts.

C. Topic-Centric Systems

Topic-centric systems, most notably wikis, organize content into collaboratively-authored articles that each focus on a single topic. A simple watchlist-rollback mechanism helps authors become aware of, and quickly repair, any damage caused by the work of other authors. Studies have shown that wiki content, despite often being contributed by non-experts, can have equivalent quality, greater currency, and much more comprehensive coverage than conventional, expert-curated sources [14]. Wikis, however, are deeply challenged by controversial topics [15][16]. They capture, by their nature, the “least-common-denominator” consensus between many authors (any non-consensus element presumably being edited out by those that do not agree with it), and the controversial core of deliberations are typically moved to massive talk pages for the article, which are essentially time-centric venues prone to all the limitations we noted above.

D. Debate-Centric Systems

Debate-centric systems, such as whysaurus.com, debatepedia.com, debatewise.org, and debate.org, have been designed to address the weakness of topic-centric systems around controversial topics. In such tools, a debate question is posed e.g. "Is the death penalty justified?", and users contribute arguments for and against that question, typically organized as two columns: one for pros, another for cons. Such tools, especially when curated to avoid duplication, provide an effective means for gathering a broad range of arguments on divisive topics, but are limited in several important ways. They are, to begin with, limited to "binary" debates where the question admits of only a "yes" or "no" answer. They are thus not suited to problems e.g. "how can we protect ourselves from climate change?" that have a large open-ended set of possible solutions. They also do not provide a systematic structure for supporting or rebutting arguments, since arguments can not be linked to other arguments. For both these reasons, the structure is not well suited for exploring open-ended deliberation problems in depth.

E. Argument-Centric Systems

Argument-centric systems [17][18] allow groups to systematically capture complex deliberations as tree structures made up of issues (questions to be answered), ideas (possible answers for a question), and arguments (statements that support or detract from an idea or argument) that define a space of possible solutions to a given problem:

Such tools have many advantages. Every unique point appears just once, radically increasing the signal-to-noise ratio, and all posts must appear under the posts they logically refer to, so all content on a given question is co-located in the tree, making it easy to find what has and has not been said on any topic, fostering more systematic and complete coverage, and counteracting balkanization by putting all competing ideas and arguments right next to each other. Careful critical thinking is encouraged, because users are implicitly encouraged to express the evidence and logic in favor of the options they prefer [19], and the community can rate each element of their arguments piece-by-piece.

Most argumentation systems have been used by individuals or in small-scale settings, relying in the latter case on a facilitator to capture the free-form interactions of a collocated group as a commonly-viewable argument map [20]. The Deliberatorium [21] is a web-based tool to allow crowd-scale online discussion and deliberation.

As we can see, argumentation systems offer much promise as a medium for enabling large-scale online deliberation. One key challenge for such systems, however, is that the logical structure of argument maps, while systematic, is not a good match with the narrative forms of knowledge communication that most people are much more familiar with. The project reported here has explored whether narratives generated *from* argument maps, using a technique called RST, can make the

 ISBN: 978-1-61208-443-5

results of argument-centric deliberations more accessible and understandable to the average user

III. RHETORICAL STRUCTURE THEORY

RST is a theory of text organization where semantically related clauses are structured hierarchically [22]. An RST structure, more specifically, is a network made up of two basic units: the nucleus and the satellite. Nuclei represent the essence of the communication, while satellites contain additional information about the nucleus. The satellite is often incomprehensible without the nucleus, whereas nuclei without satellites can be understood to a certain extent.

RST relations are classified according to their expected effect on the reader. Mann and Thompson [22] originally proposed 24 semantic relations, including: Attribution, Cause, Circumstance, Contrast, Elaboration, Enablement and Solutionhood

Figure 1 presents a sample of a RST schema. In this figure, the central information is "I use sun protection SPF30". "to prevent skin cancer" is an enablement provided by the central idea. Additionally, the entire utterance is attributed to "My mother always says".

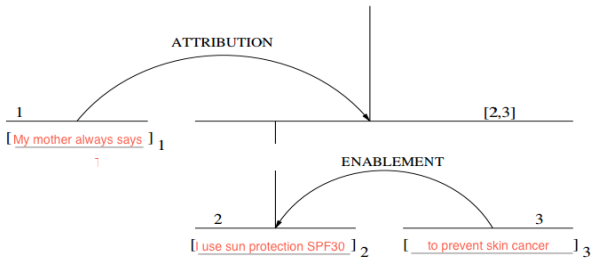


Figure 1. RST example for the text: "My mother always says that I should use sun protection SPF30 in order to prevent skin cancer".

RST has been successfully applied to many different areas over the past 30 years, and it remains a research baseline for text analysis, parsing, summarization, essay scoring and natural language generation.

"RST is a theory of text organization resulting from exhaustive analysis of texts" [23]. It was meant to propose a guideline to computational text generation based on constructs that reflect how written text works. Every piece of text is included for a perceivable reason, dictated by the RST framework, leading to a coherent text that will foster readers' understanding.

The emphasis of our research is to propose a method for building coherent text over a discussion to improve newcomers' understanding. RST offers an interesting framework to organize large online discussion. Before building an automatic tool to generate RST-based explanation over a discussion, we developed an experiment to test its effectiveness on average users' understanding.

IV. EXPERIMENTAL EVALUATION

This section describes the experiment developed to confirm our hypothesis that RST-organized explanations positively impact understanding.

A. Subjects

An invitation email was sent to 70 people associated to the computer science department of a Brazilian University. Forty-eight people accepted to participate. Individuals were randomly assigned to each group, but considering their gender, age and education level to create three demographically homogeneous groups of sixteen people each, as illustrated in Table I. Participants were mostly young (between 20 and 30 years old), male and educated. Most were computer science undergrad and graduate students. They all had substantial experience with social media tools, as well as some previous experience with forums and argument maps.

The strong participation of young people suggests the need for further studies to test the generalizability of our conclusions.

B. The Task

The task consisted of reading a debate concerning the design of a virtual coin for a new computer game, and then answering questions about it.

We initially selected a hot discussion topic that was going on in the news for quite a long time. The discussion was in an open form in the Internet that attracted many posts from many different people, concerning the Brazilian post office efficiency. There was a corruption scandal and people were discussing the need to have a public post office.

The second experiment, reported in this paper, included a competitive ingredient as the incentive to participation and a topic that participants did not have a prior opinion. Among the options considered, we decided for a discussion presented in game designers' forums.

TABLE I. PRTICIPANTS' PROFILE DESCRIPTION OF THE THREE GROUPS PARTICIPATING IN THE EXPERIMENTS.

Group	Age (years old)				Gender		Education		
	20-30	30-40	40-50	>50	Female	Male	Undergrad Student	College	MSc/PhD
Group 1 (G1)	10	4	1	1	4	12	6	7	3
Group 2 (G2)	9	4	1	2	5	11	4	10	2
Group 3 (G3)	10	4	1	1	6	10	5	9	2

 ISBN: 978-1-61208-443-5

The discussion material concerned the design of a virtual coin for a new computer game and was taken from a known website forum. The game was inspired by the Lord of Rings tale with dwarves, elves, orcs, hobglobins, and drows societies. The debate was held in a computer game community forum [24] and lasted a few days in 2009. We selected this material for the high number of posts, 69 posts generated by 36 people in a 3-days discussion.

The task consisted in reading the material received in one of the three possible formats and answering a questionnaire of 13 questions. The material in all three scenarios were displayed in an online website. In the forum scenario, user interaction was constrained to search for words, scroll the document and copy&paste material from the document to the answer slot in the questionnaire area. Furthermore, RST and ArgMap scenarios allowed obtaining additional information by clicking in the paragraph and by clicking on the node, respectively. Detailed information concerned the author and the time stamp of the posts.

The task was performed in a controlled room with 10 computers. The experiment responsible received the participants, placed them in the computer stations and red the instructions aloud. Participants asked many questions, such as if the duration was a hard constraint, if they had to write complete sentences as answers, if they could copy parts of the original material and paste as answers and what would happen with the prize in case of ties. The experiment responsible answered the questions and stayed at the control room during the entire experiment.

The experiment website first displayed the experiment's instruction. This screen contained the same material read by the experiment responsible. After reading and agreeing in participating, a second screen appears with the material displayed according to the three possible scenarios. After reading the material they would click in the OK button to start answering the questionnaire. They could go back at any time to review the material when answering the questions.

No communication was allowed. We believe they obeyed this rule because they were competing among themselves. There was a small monetary payment for participating and a prize for the best three scores.

The material, in all three scenarios, was divided in chunks of information that received a number. These numbers worked as indexes to content, working as a set of discrete options from which users could select and assemble to compose an answer. Participants could answer questions using their own words, copying and pasting sentences from the original material or by writing the "chunks of information" indexing numbers.

The experiment lasted about 1 hour. After reviewing the material, participants could start the question and answer phase of the experiment. Questions were presented one at a time in a random order. After submitting an answer, a new question was presented. Participants could go back, at anytime, to the reading material, but not to a previously answered question.

There was an incentive for participants answering correctly. The best three scores would receive a monetary prize during a later workshop, so recognition from the community was also a reward.

The questionnaire had an answers' sheet prepared by a group of two graduate students and revised by one Linguist. Most questions had just one correct answer that contained from one to 10 segments of information. Precision was calculated as the relation between the number of corrected segments in the answer and the number of segments in the answer. Recall was calculated as the relation between the number of corrected segments in the answer and the number of segments in the expected answer. We also consider the F-measure metric because it is a balance between precision and recall metrics. F-measure is the harmonic mean Precision and Recall metrics.

C. Question Types Used in Study

We focused our research on generating answers to six frequent types of questions a newcomer might have concerning a discussion: What, Compare, Explain, Justify, Choose and Summarize. An answer is designed, as shown in Table II, according to the type of the question and the expected completeness of the answer. Optional information concerning social and temporal context can also be derived.

 ISBN: 978-1-61208-443-5

TABLE II. A SAMPLE OF RHETORICAL RELATIONS FOR GENERATING ANSWERS TO DIFFERENT TYPES OF QUESTIONS.

Question Type	Question's quantifier			Context	
	One	Some	All	Chronological Factor	Social Factor
What	Purpose	Sequence	Sequence	Motivation	Motivation
Compare	---	Contrast	Contrast	Antithesis	Antithesis
Explain	Interpretation & Evaluation; Relations of Cause	Interpretation & Evaluation; Relations of Cause	Interpretation & Evaluation; Relations of Cause	Enablement	Enablement
Justify	Condition &Otherwise	Condition &Otherwise	Condition &Otherwise	Justify	Justify
Choose	Purpose	Purpose	Purpose	Evidence	Evidence
Summarize	Restatement & Summary	Restatement & Summary	Restatement & Summary	Background	Background

The rhetoric structure guides the construction rules. There are explicit rules for generating the rhetoric answers, as described below.

1. Query←Get query from user
2. QueryType← Classify(Query)
3. QueryQuantifier←ClassifyQty(Query)
4. DecomposeQuery (QueryType, Query, QueryOrganization)
5. GetAnswerComponents(QueryQty, QueryOrganization, AnswerOrganization)
6. GenerateAnswer(Answer, AnswerOrganization)

For example, suppose a question over a discussion, as an argumentation map, concerning options for buying a car, such as, How does a Toyota Rav 4 compare to a BMW X1?

1. Query← Compare a ToyotaRav4 to a BMW_X1
2. QueryType←Contrast
3. QueryQuantifier←ALL
4. QueryOrganization← (COMPARE (EXPLAIN ToyotaRav4) (EXPLAIN BMW_X1))
5. AnswerOrganization←
   (ANTITHESIS (ISSUE "Car Buying Options")
   ((ALTERNATIVE ToyotaRav4)
   (ADVANTAGE (CRITERION "Quality" "Deluxe")
   (CRITERION "Safety" "well-trusted breaks on snow"))
   (DISADVANTAGE (CRITERION "Cost" "Very high")))
   ((ALTERNATIVE BMW_X1)
   (ADVANTAGE (CRITERION "Quality" "Cool")
   (CRITERION "Quality" "Beautiful"))
   (DISADVANTAGE (CRITERION "Cost" "Very high"))))
6. Answer shown in Table III.

TABLE III. AN EXAMPLE OF A RST GENERATED ANSWER. » MEANS LINK TO INFORMATION CONCERNING AUTHOR, DATE AND SUPPORTERS.

	Car Buying Options »			
	ToyotaRav4 »		BMW_X1 »	
Criterion	Pros	Cons	Pro	Cons
Quality	Deluxe »		Cool »	
			Beautiful »	
Safety	well-trusted breaks on snow »			
Cost		Very high »		Very high »

D. Material and Apparatus

The task consisted of reading a debate concerning the design of a virtual coin for a new computer game.

Participants were divided into three groups. Each group received the reading material in one of the three formats:

- Forum format (scenario 1), as shown in Figure 2: a sequence of textual posts with date stamps and a nickname signatures.

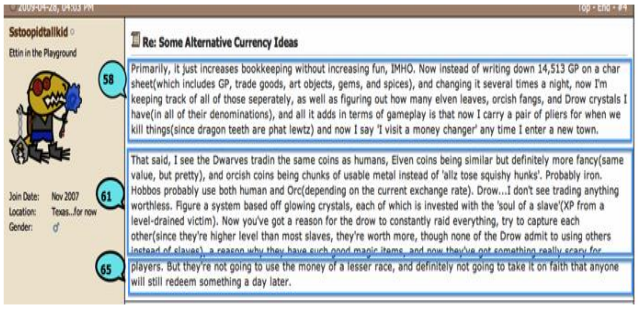


Figure 2. Discussion sample presented in a Forum format. The blue balloons represent the "chunk of information" indexing number.

- Argument map (ArgMap--scenario 2), as shown in Figure 3: the discussion from the original forum was reread and logically organized into issues, ideas, and arguments. We used the Deliberatorium tool [4] to build the argument map and the same wording as used in the original forum.

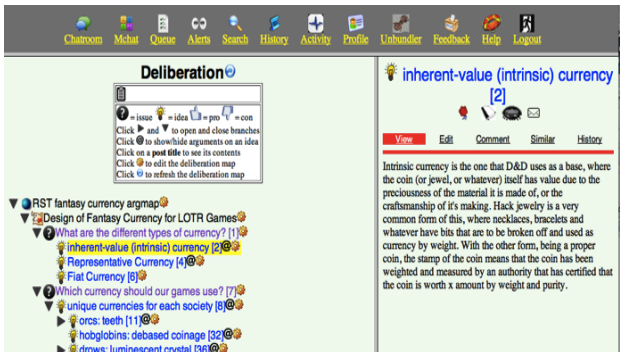


Figure 3. Discussion sample presented as an argument map. Numbers within [] represent the "chunk of information" indexing number.

- RST text (scenario 3), as shown in Figure 4: a rhetoric text that combines, temporal, logical and social aspect of the discussion, as proposed in our research.

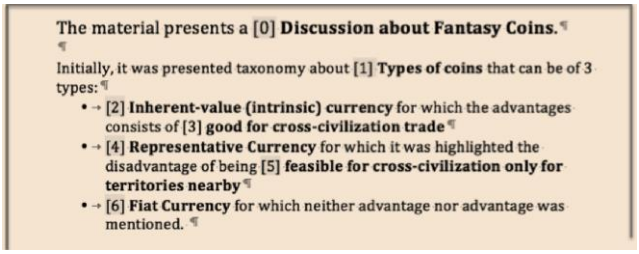


Figure 4. Discussion sample presented as a rhetoric text. Numbers within [] represent the "chunk of information" indexing number.

After reading the material, participants had to answer a questionnaire. As shown in Table IV, the questions were classified according to Bloom's taxonomy [25], reflecting the cognitive skills expected to be triggered in the participant when answering a question, such as:

- Remembering: retrieving facts;
- Understanding: interpreting the meaning of facts by being able to Exemplify, Classify, Summarize, Infer, Compare and/or Explain;
- Analyzing: breaking content into parts and detecting how the parts relate to each other and to an overall structure or purpose, being able to differentiate and organize the answer;
- Evaluating: making judgment based on criteria;

Additionally, also presented in Table IV, we decoded a possible understanding of the question using a graph database query language. Participants only received the textual questions. The corresponding computational queries were developed to explain the difficulties users might have when answering questions, as if they were computer agents. The graph representation helped to visualize the indirection of the search when answering the questions. The objective of the two last columns of Table IV was to provide readers a notion of the complexity of answering questions. The query language representation suggests the cognitive activities and efforts an agent is required to perform to answer a question, as if the material was represented in a database. The graph representation is another approach to assess the cognitive effort. The greater the number of nodes and indirections, the more complex will be to answer the question.

Both pieces of information were used to objectively measure the question's complexity and check the correlation with precision and recall metrics.

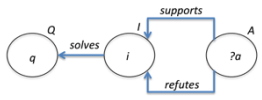
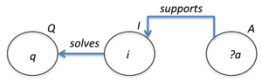
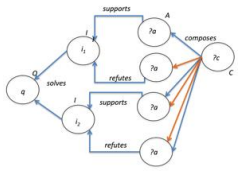
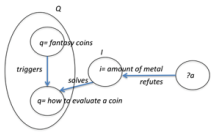
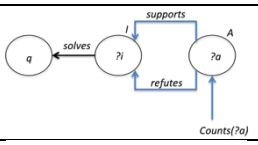
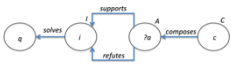
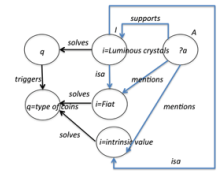
E. The Procedure

The experiment took place in November of 2013, involving 48 participants, spending about 1 hour to answer questions concerning a previous discussion held by a computer game community that lasted a few days in 2009 [24]. Their task was to read the material concerning the discussion and to answer 13 questions about it. All the participants were assured that their information would remain anonymous.

The experiment took place in a controlled room with one computer per participant. A moderator read the experiment description, the permit form and the instructions. Participants were told they could quit at anytime. Actually, two participants quitted. The moderator remained in the room until the end of the experiment. There was no communication allowed among participants during the experiment started.

 ISBN: 978-1-61208-443-5

TABLE IV. QUESTIONNAIRE SAMPLE WITH A QUESTION PROPERLY CLASSIFIED.

Id	Bloom's Classification	Question	Analogous computational question	Graph Representation
1	Remember	What are all arguments for using "Teeth as coins" for the "Orchs civilization"?	$answer(?x,y) \leftarrow (?x\ isa\ Argument), (y\ isa\ Idea),$ $(y =$ "Teeth as coins "$),$ $(?x\ supports\ y)\|(?x\ refutes\ y)$ ¶	
2	Remember	Provide, at least 2 positive evidences, to use "Dungeon &Dragon (D&D) coins with exotic names for each civilization"?	$answer(?x,y) \leftarrow (?x\ isa\ Argument), (y\ isa\ Idea),$ $(y =$ "Dungeon & Dragon"$), (?x\ supports\ y),$ $count(?x) \geq 2$ ¶	
3	Understand	What are the similarities and differences of using "Teeth for the Orch civilization" and "Luminous Crystals for the Drows"? Provide at least two of each	$answer(\pi_1,\pi_2) \leftarrow (c\ isa\ Criterion),$ $(i_1\ isa\ Idea), (i_2\ isa\ Idea),$ $(i_1 =$ "$Teeth\ for\ Orch$"$),$ $(i_2 =$ "Luminous Crystals for Drows"$),$ $(i_1\ \pi_1 c), (c\ \pi_2 i_2)$ ¶ where π_1 and π_2 are the actual paths that matched	
4	Understand	Provide an argument that weakens the option of using the amount of metal within a coin as a way to estimate the value of a coin	$answer(?a,i) \leftarrow (?a\ isa\ Argument),$ $(i\ isa\ Idea),$ $(q_1\ isa\ Question), (q_2 is\ a\ Question), (q_2 triggers\ q_1),$ $(i\ isa\ Argument), (q_1 =$ Fantasy coins$),$ $(q_2 =$ "How to estimate the value of a coin"$),$ $(i\ solves\ q_2), (?a\ refutes\ i)$ ¶	
5	Evaluate*	Which civilization produces coins with the best quality metal?	$asnwer\left(civilization, best\left(evaluate(a)\right)\right)$ $\leftarrow (i\ isa\ Idea), (a\ isa\ Argument), (c\ isa\ Criterion),$ $(i\ mentions\ civilization), (a\ supports\|refutes\ i),$ $(c\ composes\ a), (c = foreign\ trading)$	
6	Evaluate *	Which coin seems the best for foreign trading?	$asnwer(i, best(evaluate(a))) \leftarrow$ $(a\ isa\ Argument), (c\ isa\ Criterion),$ $(a\ supports\|refutes\ i), (c\ composes\ a), (c =$ $foreign\ trading)$	
7	Evaluate *	Which coin was most discussed?	$answer\ (?i, \max\ (count(a)) \leftarrow$ $\leftarrow (?i\ isa\ Idea), (a\ isa\ Argument), (a\ supports\ i)$	
8	Understand	What was said about "Luminous Crystals" for "foreign trading"?	$answer(?a,c) \leftarrow (?i\ isa\ Idea),$ $(c\ isa\ Criterin),$ $\left((?a\ supports\ i)\middle\|(?a\ refutes\ i)\right),$ $(c\ composes\ ?a), (i =$ "Luminous Crystals"$),$ $(c =$ "$Foreign\ trading$"$)$	
9	Remember	What is the complete list of coins proposed for the Dwarves civilization?	$answer(?i,q) \leftarrow (?i\ isa\ Idea), (q\ isa\ Question),$ $(q =$ "Fantasy coins"$),$ $(?i\ solves\ q),$ $(?i\ mention$ "$Dwarves$"$))$	
10	Evaluate	Do you think, according to the text, that the "Luminous Crystals" are best classified as a fiat or as an intrinsic value coin? Explain your choice	$answer(?i, best((i_1, evaluate(?a_1)), (i_2, evaluate(a_2))))$ $\leftarrow (?i\ isa\ Idea), (q_1\ is\ a\ Question), (q_2 isa\ Question),$ $(q_1\ triggers\ q_2), (q_1 =$ Fantasy Coins$), (q_2 =$ Type of Coins$),$ $(i_1 isa\ Idea), (i_2 isa\ Idea), (i\ isa\ Idea), (i_1 = Fiat), \ (i_2$ $= Intrinsic\ Value), \ (i$ $= Luminous\ Crystals), (?a_1\ supports\ i), (?a_1 mentions\ i_1),$ $(?a_2 supports\ i), (?a_2 mentions\ i_2)$	
11	Analyze	What are all arguments supporting "Teeth as coins" related to the "intrinsic value" of a coin?	$answer(x,y) \leftarrow (x\ isa\ Argument), (y\ isa\ Idea),$ $(c\ isa\ Criterion), (x\ supports\ y), (c\ composes\ x),$ $(y =$ "Teeth as coins"$), (c =$ "Intrinsic value"$)$	
12	Remember	What is the main discussion all about?	$answer(x) \leftarrow (x\ isa\ Question), (x\ Mention$ "Main"$)$	
13	Analyze	What are all arguments supporting the claim that Dwarves coins are good for foreign trading?	$answer(?a,c) \leftarrow (q\ isa\ Question), (?i\ isa\ Idea),$ $(?a\ isa\ Argument), (c\ isa\ Criterion),$ $(?i\ solves\ q), (?i\ mentions$ "$Dwarves$"$),$ $(?a\ supports\ ?i), (c\ composes\ ?a),$ $(q =$ Fantasy Coins$), (c =$ "$Foreign\ trading$"$)$	

F. The Metrics

We considered a set of 26 variables, organized in four groups, as described in Table V, to select the statistically significant ones that might affect the results.

TABLE V. THE SET OF INDEPENDENT VARIABLES.

Variable Type	Variable ID	Description
Material	MT	Forum, argumentation map or RST text
	MNumLetters	Number of letters in the displayed material
	MWC	Number of words in the displayed material
	MBC	Number of blocks in the material. Blocks are posts in forum, nodes in argumentation maps and paragraphs in RST text.
	MIdentation	Maximum indentation of displayed material
Question	QT	Question type according to Bloom's taxonomy
	QNumLetters	Number of letters in the question
	QWC	Number of words in the question
	QQ	Question quantifier: all, some or one
	QClauses	Question's number of clauses
	QNodes	Question's number of nodes
	LLinks	Question's number of links
Expected Answer	EANodes	Number of nodes in the expected answer
	EAFirstNode	Smallest node number in the expected answer
	EALastNode	The greatest node number in the expected answer
	EAMaxSpam	EASpam=EALastNode - EAFirstNode
Participant 's Answer	PANodes	Number of nodes in the participant's answer
	PAFirstNode	The smallest node number in the answer
	PALastNode	The greatest node number in the answer
	PAMaxSpam	PASpam=PALastNode - PAFirstNode
	PADFirstLast	Number of letter from PAFirstNode to PALastNode
	PANumLetters	Number of letters in the participant's answer
	PAWC	Number of words in the participant's answer
	PACNodes	Number of corrected nodes in the participant's answer
	TRM	Participant's time to read the material
	TUAQ	Participant's time to answer a question

The dependent variable included the classic metrics of document retrieval domain, as described in Table VI.

TABLE VI. THE SET OF DEPENDENT VARIABLES.

Variable Type	Variable name	Description
Participant's Answer	Precision	$\frac{PACNodes}{EANodes}$
	Recall	$\frac{PANodes}{EANodes}$
	F-Measure	$\frac{(2 * Precision * Recall)}{(Precision + Recall)}$
	PrecisionHit	$If\ Precision = 1,$ $Then\ PrecisionHit = 1$ $Else\ PrecisionHit = 0$
	RecallHit	$If\ Recall = 1,$ $Then\ RecallHit = 1$ $Else\ RecallHit = 0$
	F-MeasureHit	$If\ F\_measure = 1,$ $Then\ F\_measureHit = 1$ $Else\ F\_measureHitHit = 0$

G. Statistical Analysis

The comparison of means test was performed for each of the 13 questions considering the F-measure metric for being a balance between precision and recall. We considered three comparison scenarios:

- Test1: RST and Forum, the null hypothesis is that the F-measure for the RST scenario is not significantly higher than in the Forum scenario;
- Test2: RST and Arg. Map, the null hypothesis is that the F-measure for the RST scenario is not significantly higher than in the Arg. Map scenario;
- Test3: Arg. Map and Forum, the null hypothesis is that the F-measure for the Arg. Map scenario is not significantly higher than in the Forum scenario.

The T-test [26] assumes that samples are randomly drawn from normally distributed populations with unknown population means. For this reason, before performing each of the t-tests, the Kolmogorov-Smirnov test [26] was performed to check the hypothesis of normality. The hypothesis of normally distribution data was only observed for questions 1, 3, 8, 11 and 13. Table VII presents p-values for the three tests. P-value reflects the probability of proving the null hypothesis, i.e., the probability that our hypothesis is false [26]. For questions that did not pass the normality distribution, it was possible to evaluate the proportion of hits for precision and recall, as shown in Table VIII.

TABLE VII. P-VALUES FOR F-MEASURE METRIC. GREEN CELLS HIGHLIGHT P-VALUE < 0.05.

Question	Test 1: RST and Forum	Test 2: RST and Arg. Map	Test 3: Arg. Map and Forum
Q1	0.004639	0.8278	0.0001833
Q3	0.01968	0.006144	0.786
Q8	0.02396	0.05142	0.3286
Q11	0.1363	0.2848	0.2274
Q13	0.1124	0.3507	0.1585

 ISBN: 978-1-61208-443-5

TABLE VIII. PRECISION AND RECALL "HIT" P-VALUES. GREEN CELLS REPRESENT P-VALUE <0,1.

	HitPrecision			HitRecall		
Question	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Q1	0.00013037	0.2326044	0.001140098	0.07206352	0.9002284	0.00745985
Q2	0.05123522	0.5	0.05123522	0.03137432	0.6868228	0.01105973
Q3	0.1439504	0.03826125	0.7673956	0.1548145	0.1548145	0.5
Q4	0.5	0.5	0.5	0.1548145	0.1548145	0.5
Q5	0.1548145	0.5	0.1548145	0.1548145	0.5	0.1548145
Q6	0.2720985	0.2720985	0.5	0.2720985	0.2720985	0.5
Q7	0.2720985	0.07206352	0.8174849	0.2720985	0.07206352	0.8174849
Q8	0.00114010	0.01625472	0.1439504	0.3131772	0.3131772	0.5
Q9	0.1548145	0.03442252	0.8574753	0.07206352	0.07206352	0.5
Q10	0.00328921	0.5	0.003289207	0.1425247	0.03442252	0.8451855
Q11	0.06356221	0.5	0.06356221	0.5	0.8451855	0.1548145
Q12	0.5	0.5	0.5	0.5	0.5	0.5
Q13	0.07206352	0.8451855	0.01625472	0.06356221	0.2326044	0.2071081

We had to discharge task duration from our analysis, since there was too much noise in this measurement because some, but not all, participants did a great deal of copying from the original material and pasting as an answer. This answering method was fair, but the time spent typing an answer and the time spent copying&pasting could dramatically mask the results.

F-measure for questions 1, 3 and 8, as well as PrecisionHit and RecallHit for questions 1, 2, 8, 9, 10, 11 and 13 support our claim that RST-organized text improves newcomers understanding of a discussion. Two questions were raised from these results: "Why were not good the results for questions 4, 5, 6, 7 and 12?" and "What may explain the good performance of RST?".

First of all, let us investigate possible reasons for the bad results. Questions 5, 6 and 7 concern providing the "best perceived value" for an entity: "civilization with best quality coin", "the best coin for foreign trade" and "the most discussed coin", respectively. Although intuitively these questions should have triggered a search, evaluation, comparison and selection processes, participants perform as in a pattern matching style. They answered very fast and their answers presented very high precision and recall values, no matter the method. The task becomes just a matter of retrieving from the text the information that caused the highest impact in the reader. Thus, no matter the method, it will be a matter of being impacted by the information. When we are asked for explaining our selection, as triggered in question 10, probably the 4-stage process is called to take place. In this later case, RST text method played an important role facilitating newcomers' understanding.

Question 12 was also a question to check the minimum attention of the participants. It was a pretty easy question with very high precision values no matter the method. Actually, we could use this question as a filter: only consider respondents with high values in this question.

We are still investigating the possible answers for the bad results of question 4

Although we had very good results for an exploratory research, we wanted to understand why the RST format was causing such positive impact. We investigated the 26 independent variables searching for single or group correlations that would explain the results on the 624 answers.

We used the LASSO statistic method [27] to select the relevant variable to run a linear regression model. The method penalizes models with higher number of variables. It does a balance between quality and complexity.

V. CONCLUSION AND FUTURE WORK

The contribution of this research was to show that the presentation format of a group discussion impacts comprehension. Furthermore, a rhetorically organized text improves understanding, especially for answering cognitively complex questions, over classic sequential forum's organization. We are currently implementing a tool developing according to our RST answer generator method. We did a small experiment to explore the idea.

The results, as presented in Table VIII, were very promising and interesting, even when p-value was not small enough to refute the null hypotheses.

From the 26 variables, results indicated HitF-measure (0 or 1 value) was mostly affected by EAMaxSpam parameter. The results indicate, with p-value<0.05, that F-measure is inversely affected by the size of the spam of the expected answer (EASpam).

Inspecting our automatic RST text generation, we realize that this is exactly what the method is meant to do: grab the relevant information pieces and organize them in a concise text, bringing together time, logic and social information to provide context to the message.

This is still a first, but promising step towards specifying the design of large-scale collaboration environments. Although we did a comparison study, we believe RST texts can be used as a storyteller, guiding participants through logical (argument map representation), temporal (forum) and social aspects of a discussion.

 ISBN: 978-1-61208-443-5

The main objective of our research was accomplished. We could show that making sense of an ongoing discussion can be sensibly improved by using RST-based textual explanation generated from argument map organized discussion. Additionally, our observations on participants' behavior answering questions about the discussion raised the possibility of integrating this Q&A feature to crowd source answers that would be generated exploring the material of a discussion done by experts.

While RST was developed as a descriptive technique for analyzing natural language text, it can also be used prescriptively to describe how logical points can be structured in order to be persuasive and clear. Given that RST structures demonstrably increase comprehensibility of complex content, our next step is to explore how we can generate RST structures automatically for real-world online discussions. Our strategy for this will include:

- generating argument maps - natively or by argument mining [28][29] from web forums
- developing algorithms to automatically generate RST-structured responses to queries from these argument maps, building upon on a taxonomy of canonical query types which each have an RST template plus rules describing how to harvest the argument map information needed to fill in the empty template slots.

REFERENCES

[1] M. A. Andersen, "Asynchronous discussion forums: success factors, outcomes, assessments, and limitations," Educational Technology & Society, vol. 12 (1), 2009, pp. 249–257.

[2] D. Feng, E. Shaw, J. Kim, and E. H. Hovy, "An intelligent discussion-bot for answering student queries in threaded discussions," Proc. 11th Intelligent User Interface Conference, 2006, pp. 171-177.

[3] J. Conklin, A. Selvin, S. B. Shum, and M. Sierhuis, "Facilitated hypertext for collective sensemaking: 15 years on from Gibis," Proc. 8th International Working Conference on the Language Action Perspective on Communication Modelling (LAP'03), July 2003, pp. 1-22.

[4] M. Klein, "The MIT Deliberatorium: Enabling Large-Scale Deliberation About Complex Systemic Problems," Proc. International Conference on Agents and Artificial Intelligence, 2011, pp. 15-24.

[5] W. Kunz and H. Rittel, "Issues as Elements of Information Systems", Working Paper 131, Center for Planning and Development Research, University of California, Berkely, CA, 1970.

[6] S. B. Shum and A. M. Selvin, "Structuring discourse for collective interpretation," Proc. Distributed Collective Practices: Conference on Collective Cognition and Memory Practices, 2000, pp. 1-16.

[7] V. Uren, S. B. Shum, G. Li and M. Bachler, "Sensemaking Tools for Understanding Research Literatures: Design, Implementation and User Evaluation," International Journal of Human Computer Studies, vol. 64(5), 2006, pp. 420–445.

[8] U. Hermjakob, "Parsing and question classification for question answering," Proc. ACL Workshop on Open-Domain Question Answering, vol. 12, 2001, pp. 1-6.

[9] F. H. Eemeren and R. Grootendorst, A Systematic Theory of Argumentation: The Pragma-dialectical Approach. Cambridge, MA: Cambridge University Press, 2003.

[10] D. N. Walton and E. C. W. Krabbe, Commitment in dialogue: Basic concepts of interpersonal reasoning. Albany, NY: State University of New York Press, 1995.

[11] M. Klein and G. Convertino, "A Roadmap for Open Innovation Systems," Journal of Social Media for Organizations, vol. 1 (2), 2015, pp. 1-16.

[12] M. Klein and G. Convertino, "An Embarrassment of Riches: A Critical Review of Open Innovation Systems," Communications of the ACM, vol. 57(11), 2014, pp. 40-42.

[13] F. Calefato, F. Lanubile, F., M. Raffaella and N. N. Merolla, "Success Factors for Effective Knowledge Sharing," Proc. 10th International Forum on Knowledge Asset Dynamics (IFKAD'15), Jun 2015, pp. 1-11.

[14] J. Giles, "Internet encyclopaedias go head to head," Nature, vol. 438(7070), 2005, pp. 900-901.

[15] A. Kittur, B. Suh, B. A. Pendleton and E. H. Chi, "He says, she says: conflict and coordination in Wikipedia," Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM Press, 2007, pp. 453-462

[16] F. B. Viegas, M. Wattenberg and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," Proc. SIGCHI conference on Human factors in computing systems, ACM Press, 2004, pp. 575–582.

[17] P. A. Kirschner, S. J. B. Shum and C. S. Carr, Visualizing Argumentation: Software tools for collaborative and educational sense-making, Springer, 2003.

[18] A. D. Moor and M. Aakhus, "Argumentation Support: From Technologies to Tools," Communications of the ACM, vol. 49(3), 2006, pp. 93-98.

[19] C. S. Carr, "Using computer supported argument visualization to teach legal argumentation," in Visualizing argumentation: software tools for collaborative and educational sense-making, P. A. Kirschner, S. J. B. Shum and C. S. Carr, Eds. Berlin: Springer-Verlag, 2003, pp. 75-96.

[20] S. J. B. Shum, A. M. Selvin, M. Sierhuis, J. Conklin and C. B. Haley, "Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC," in Rationale Management in Software Engineering, A. H. Dutoit, R. McCall, I. Mistrik and B. Paech, Eds. Berlin: Springer-Verlag, 2006, pp. 111-132.

[21] M. Klein, "Enabling Large-Scale Deliberation Using Attention-Mediation Metrics," Computer-Supported Collaborative Work, vol. 21(4), 2011, pp. 449-473.

[22] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: towards a functional theory of text organization". Text, 8 (3), 1988, pp. 243-281.

[23] M. Taboada and W. C. Mann, Rhetorical Structure Theory: looking back and moving ahead, Discourse Studies, London: SAGE, 2006, pp. 423-459.

[24] Giant in the Playground Forum: http://www.giantitp.com/forums/showthread.php?110342-Some-Alternative-Currency-Ideas. Accessed September, 8th, 2015.

[25] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives, New York: Pearson, Allyn & Bacon, 2001.

[26] R. S. Witte and J. S. Witte, Statistics, Wiley, 10th edition, 2013.

[27] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, 58(1), 1996, pp. 267-288.

[28] B. Pang and L. Lee, "Opinion mining and sentiment analysis". Foundations and trends in information retrieval, vol 2(1-2):1-135, (2008) .

[29] C. Reed and G. Rowe, "Araucaria: software for argument analysis, diagramming and representation," International Journal of Artificial Intelligenc Tools, vol. 13(4), 2004, pp. 961-979.

 ISBN: 978-1-61208-443-5