A Visual Interface for Analyzing Text Conversations

Shama Rashid[1] and Giuseppe Carenini[2]

[1] University of British Columbia, Vancouver, BC, Canada,
shama.rashid@gmail.com
[2] University of British Columbia, Vancouver, BC, Canada,
carenini@cs.ubc.ca

Abstract. This paper presents a visual, intelligent interface intended to help user analyze possibly long and complex conversations. So far, the interface can only deal with synchronous conversations, but most of its components could be also applied to asynchronous conversations such as blogs and emails. In a Business Intelligent scenario, our interface could support the analysis of real-time conversation occurring for instance in blogs, discussion fora, or in sites intended to collect customer feedback. The design of our interface aims to effectively combine the power of human perceptual and cognitive skills with the ability of Natural Language Processing (NLP) techniques to mine and summarize text. Since we are targeting a large user population, with no, or minimal expertise in data analysis, we selected interface elements based on simple and common visual metaphors. Furthermore, to accommodate for user differences in such a large population, we provided users with the ability to satisfy the same information needs in different ways.
We have tested our interface in a formative user study, in which participants used the interface to analyze four long and complex conversations, in order to answer questions addressing specific information needs in a business scenario. This evaluation revealed that the interface is intuitive, easy to use and provides the tools necessary for the task. Participants also found all the interface components quite useful, with the main problems coming from inaccuracies in the information extraction process, as well as from deficiencies in the generated summaries. Finally, it seems that the choice of offering redundant functionalities was beneficial. The logged interaction behaviors reveal that different users actually selected very different strategies, even independently from their performance. And this was consistent with the high variability in the user preferences for the different interface components.

Key words: interacive interface, multi-modal interface, ontology, conversation visualization, conversation browsing and summarization

1 Introduction

In our daily lives, we have conversations with other people in many different modalities. We email for business and personal purposes, attend meetings in

person and remotely, chat online, and participate in blog or forum discussions. The Web has significantly increased the volume and the complexity of the conversational data generated through our day to day communication and has provided us with a rich source of readily available private and public discourse data.

It is clear that automatic summarization can be of benefit in dealing with this overwhelming amount of information by providing quick access to possibly large and complex conversations that are constantly evolving in real-time. Automatic meeting summary would allow us to prepare for an upcoming meeting or review the decisions of a previous group. Email summaries would aid corporate memory and provide efficient indices into large mail folders. Summaries of blogs could help people join large ongoing conversations in real-time and support them in searching and browsing particular sub-topics and discussions.

Previous work on summarizing conversations, however, indicates that summarization techniques generating generic, exclusively textual summary are often insufficient [11]. In contrast, on the one hand, [7] shows that meeting summaries focused on user's information needs are more effective than generic ones, even if they contain several inaccuracies. On the other hand, [18] demonstrates that visualization and flexible interactive techniques can compensate for the limitations of state-of-the-art summarization methods, especially when applied to complex text mining tasks.

Building on these findings, in this paper we present a novel interface that takes advantage of human perceptual and cognitive skills in conjunction with automatically extracted information and summarization capabilities to support users in analyzing a conversation, to satisfy a set of related information needs. The automatically extracted information includes an annotation of each utterance in the conversation for what entities it is referring to and also for the dialog acts it is expressing, such as whether it is conveying a positive vs. negative subjective orientation, a decision, a problem, or an action item.

In addition to aiming for an effective symbiosis between human perceptual and cognitive skills and the computer ability to mine and summarize text, the design of our interface follows three basic principles.

- Simplicity: the interface components are based on common visual metaphors (facets [21] and word clouds [1]), so that the interface can be used by a large user population.
- Redundancy: most common information needs can be satisfied in multiple ways, by performing different sequences of visual and interactive actions. The goal is to accommodate for, as well as to investigate, user differences in preference and expertise, that we expect to be present in such a large population.
- Generality: although we have only applied our interface to meeting transcripts, most of the underlying algorithms for information extraction and the information visualization techniques are not meeting specific and would work on different conversational modalities (e.g., emails, blogs).

We have tested our interface in a formative user study, in which participants used the interface to analyze four long and complex conversations, in order to answer questions addressing specific information needs. This evaluation revealed

that the interface is intuitive, easy to use and provides the tools necessary for the task. Participants also found all the interface components quite useful, with the main problems coming from inaccuracies in the information extraction process, and from deficiencies in the generated summaries. In terms of interactive behavior, despite a rather substantial training, participants exhibited a broad range of strategies, with no clear cut distinction between top and bottom performers.

As a preview, in the remainder of the paper, we first describe the NLP techniques for information extraction and summarization underlying our approach. Next, we present the design and implementation of the interface. After that, we discuss related work, and conclude by presenting the user study and discussing the key findings.

2 NLP: Information Extraction and Focused Summarization

Our browsing and summarization methods rely on mapping the sentences in a conversation to an ontology containing three core upper-level classes for participants (i.e., speakers), dialogue acts (DAs) (e.g., decision), and referred entities (e.g., battery). An example for mapping a sentence to the ontology is shown in Figure 1.

A: *Let's go with a simple chip.*
Speaker: A, who is the Project Manager
Entities: simple chip (only one for this particular example)
Dialog Acts: classified as decision and positive-subjective

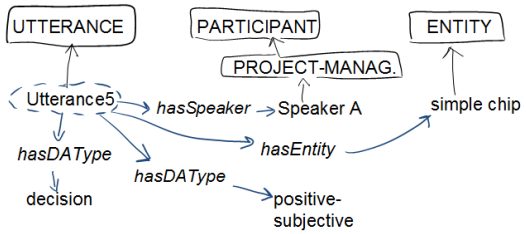

Fig. 1. Example of mapping a sentence to an ontology

We represent our ontology in OWL/RDF (Web Ontology Language/Resource Description Framework) [11]. The mapping shown in Figure 1 would be expressed in OWL/RDF as follows:

```
<Utterance rdf:about="#TS3012a.A.dialog-act.vkaraisk.14">
   <rdf:type rdf:resource="&owl;Thing"/>
   <hasSpeaker rdf:resource="#ProjectManager"/>
   <hasDAType rdf:resource="#Decision"/>
```

```
    <hasDAType rdf:resource="#PositiveSubjective"/>
    <begTime>18.61</begTime>
    <endTime>20.49</endTime>
</Utterance>
```

In term of processing, the Participant annotations for the sentences are extracted directly from the transcript of the conversation. As for the Entity class, it contains noun phrases referred to in the conversation with mid-range (10%-90%) document frequency after filtering for non-content words and stop words (words like 'anyone', 'okay' etc.). This selection strategy avoids considering overly general/specific terms. The DA annotations for sentences are determined using supervised classifiers. Our classifiers are designed for identifying five subclasses of the DA-type class, namely action items, decisions, negative and positive subjective, and problems. However, we could easily include additional classifiers to identify other types of DAs to support a larger variety of the information needs. The classifiers rely on a feature set related to generic conversational structure [**?**], including features like sentence position in the conversation and in the current turn, pause-style features, lexical cohesion, centroid scores, and features that measure how terms cluster between conversation participants and conversation turns. The classifiers also use sentence level features like word pairs, Part of Speech (POS) pairs, character trigrams etc. Overall the classification accuracy is quite high ranging from .92 to .76 (AUROC metric). Notice that these classifiers do not rely on any feature specific to conversations in a particular modality (e.g., meetings), so they can be applied to conversations in other modalities and even to multi-modal conversations (e.g., a conversation started in a meeting and continued via email). This follows our design principle of generality.

The automatic mapping of the sentences into the ontology is the key knowledge source for generating focused summaries that satisfy some given user information needs. These summaries can be either extractive or abstractive (see redundancy principle). An extractive summary is simply generated by extracting a subset of the sentences in the conversation satisfying those needs, while an abstractive summary is generated by extracting and aggregating information satisfying those needs and then generating new sentences expressing the selected information.

We generate abstractive summaries by following the approach originally proposed in [10, 11]. First, sentences are aggregated into messages based on whether, for instance, they are uttered by the same participant and express the same dialog act on the same entity. Then, the most informative messages are selected by using an optimization function combining sentences and messages subject to three constraints: a summary length constraint and two constraints tying the information value of messages and sentences together. A Natural Language Generation component is finally used to produce summary sentences from the selected messages.

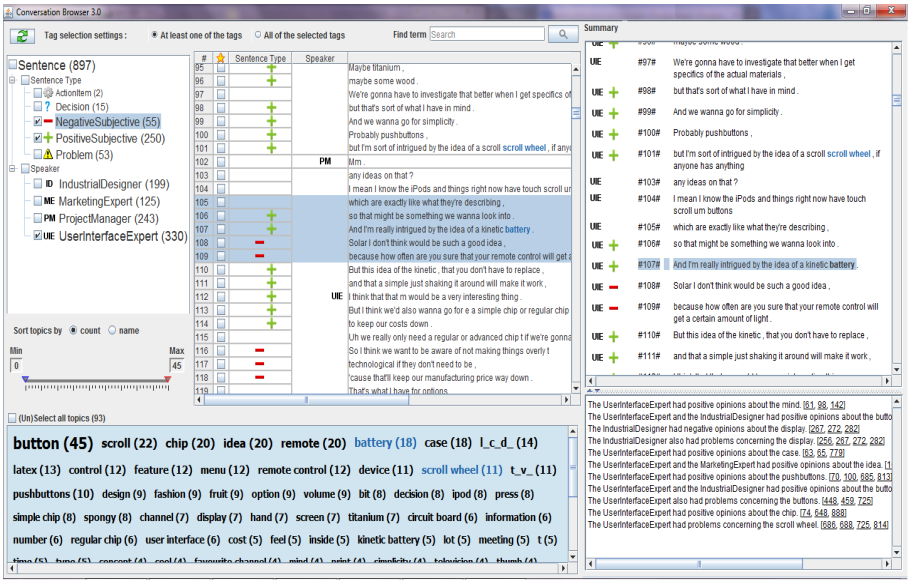

Fig. 2. A visual interface for analyzing conversations with 4 integrated views: a) the Ontology View (top left), b) the Transcript View (middle), c) the Entity View (bottom left), and d) the Summary View (right)

3 Interface Design

3.1 Display Design

As shown in Figure 2 our visual interface consists of four integrated views, the Ontology View (upper left), the Entity View (bottom left), the Transcript View (middle) and the Summary View (right). They all contribute, in an often redundant way, to support the exploration and analysis of the conversation through its mapping of sentences into an ontology, as well as through the ability of generating focused summaries that cover only aspects of the conversation that are especially relevant to the user's information needs.

The Ontology View The Ontology View provides a systematic way for the users to explore the relevant concepts in the conversation and their relations. It contains a faceted tree hierarchy with core nodes Speaker and DA-type. The root node in the ontology tree represents all the sentences in the conversation, while any other node represents a subset or subclass of those sentences that satisfy a particular property. For instance, the node ProjectManager (PM) represented all the sentences uttered by the PM, while the node ActionItem represented all the utterances that were classified as containing an action item. For leaf nodes, the counts beside the labels act as information scent and indicate how many sentences are mapped to this node. For instance, in the Speaker subtree these counts provide a sense of how dominant a speaker was in this particular

conversation. Also, notice in Figure 2 that leaf labels are scaled according to their count, to make more frequent nodes stand out more.

The user can select multiple nodes on the Ontology View tree using checkboxes juxtaposed to the node labels. To take advantage of visual popout, we have associated an icon with each of the ontology concepts. Following domain convention, we are using a green '+' or a red '-' shaped icon for PositiveSubjective and NegativeSubjective nodes respectively. For the other DA-type nodes, we have used the shape of the most common icon found when we googled using the keywords. For speaker nodes, abbreviations of the speaker names are used. For instance, the icon for the Industrial Designer is ID.

The Entity View The Entity View is a textual collage of the list of entities mentioned in the conversation represented as a tag cloud (with actual counts in parenthesis). This view not only represents a quick overview of the content of the whole conversation, but also provides access points into the conversation. The user can search the conversation for sentences mentioning any subset of the displayed entities by simply selecting that subset.

The Transcript View The Transcript View shows the sentences of the conversation one per row, ordered temporally in the 'Sentence' column. Gridlines are included to make the separation of the sentences apparent. However, sentences belonging to a turn by the same speaker are grouped using containment within a larger grid box. Notice that if an entity has been selected in the Entity View, it will be highlighted in the transcript. Moving now to the left of the sentence column (see Figure 2), additional information is shown in auxiliary columns. The two columns,'Sentence Type' and 'Speaker', show, for each sentence, the icons for the DA-type and Speaker to which the sentence was mapped. The next column to the left (with header star icon) allows the user to mark a sentence as important for her current information needs, while the leftmost column numbers utterances in the order in which they were uttered.

The Summary View The Summary View comprises two panels. The top panel displays a focused extractive summary, while the bottom one displays a focused abstractive summary. The extractive summary includes all the utterances that were mapped to concepts on the Ontology View and the Entity View currently selected by the user. Notice that the user can choose whether she wants to include utterances mapped in at least one of the concepts vs. utterances mapped to all the selected concepts. For instance in Figure 2, the user made the former choice, so the extractive summary includes all the utterances that are either subjective (positive or negative), or uttered by the Industrial Designer, or containing at least one of the selected entities (i.e., battery and scroll wheel). While the extractive summary is simply a subset of the utterances in the conversation transcript, the abstractive summary in the bottom panel comprises sentences that are generated from scratch as described in the Information Extraction and Focused Summarization Section. The abstract sentences summarize

all the utterances that in the corresponding extractive summary are simply included verbatim. Finally, as shown in Figure 2 (right bottom), each abstractive sentence is followed by a list of line numbers that link the abstractive sentence to conversation utterances the sentence is summarizing.

3.2 Interactive Techniques

When the Transcript View is generated for a conversation, the 'Sentence Type' column is initially empty and all the nodes on the ontology tree in the Ontology View are shown fully expanded and are de-selected. The user can minimize(/expand) nodes she is-not(/is) interested in on this tree hierarchy.

Once the user selects a node (or de-selects an already selected node) on the ontology tree, the keyword or icon associated with that node appears in (or disappears from) the 'Sentence Type' column of all the rows that contain sentences that can be mapped to that particular node; in case of Speaker nodes, the icons are shown all the time to keep the user oriented to who is saying what. However, selecting a Speaker type node on the Ontology View changes the filtering criteria for the summary. Once the user has selected the nodes of interest from the ontology tree and the Entity View tag cloud, she can scroll through the Transcript View and inspect sentences that appear to be promising to satisfy her information needs.

The Tag Selection Settings control (top left) allows the user to choose whether she wants to include utterances mapped in at least one of the selected concepts vs. utterances mapped to all the selected concepts.

A Range Slider is provided for the Entity View (on the top left corner of the view) to narrow down the set of displayed entities based on how many times they occurred in the conversation. Entities outside the selected range fade out.

The Entity Sort Order control (on top of the Range Slider) can be used to display the entities in the Entity View in alphabetical order or according to their frequency counts.

The Summary View is linked to the Transcript View. When the user clicks on a sentence in the extractive summary or on a link in the abstractive summary, the corresponding sentence in the Transcript View is highlighted, along with the two preceding sentences and the two subsequent sentences to make the highlight easier to spot. Simultaneously, the viewport on the Transcript View is adjusted to show the highlighting by auto-scrolling.

Finally, the interface provides a standard Keyword Search Box (top middle) and a Reset button. The Search Box allows the user to inspect the sentences in the transcript referring to a search term, one at a time, until the end of the transcript has been reached. The Reset Interface button at the top left corner of the interface allows the user to deselect all currently selected nodes for speakers, DAs and entities.

3.3 Implementation

Our prototype has been developed iteratively, using Java Swing and AWT components and Jena, an open source Java framework that provides a programmatic environment for building semantic web applications. The NLP techniques for information extraction and summarization were implemented in Python. We have also used Python scripts to do the data processing at the back end of the interface.

4 Related Work

Our approach to support the exploration of conversations by explicitly showing a mapping of all the utterances in an ontology is an example of the highly successful and popular faceted interface model [13]. Flamenco [21] and Mambo [6] are two influential systems based on facets. They make use of hierarchical faceted metadata for browsing through image or music collections, respectively. As it is commonly done in faceted interfaces, we have included a count beside each node of the ontology to indicate the number of utterances in the conversation that have been mapped to it.

The distribution of terms in a text document has been explored using different techniques (e.g., tag cloud [1], TextPool [3], Wordles [17]) ; the central theme of all of them being the use of size to encode the frequency of occurrence of a term. We have adopted a tag cloud format to list the entities in our interface, because of its simplicity and popularity.

Several aspects of our approach have been investigated in previous work on supporting the analysis of meeting conversations. The Ferret Meeting Browser [19] allows navigation by clicking on a vertical scrollable timeline of the transcript. More recently, the Meeting Miner [2] enhances timeline-based navigation, with the possibility of retrieving a set of speech turns by specifying keywords that can be typed in or selected from a list of automatically generated keywords and topics. Similarly, we provide these functionality in the Entity View and in the Keyword Search Box.

The summarization system presented in [7] makes use of dialog acts for generating an extractive decision-focused summary suitable for debriefing tasks for conversations about designing a new product (see AMI corpus [4]); our approach also considers the speaker, subjectivity and entity information. Furthermore, it provides flexibility in choosing the summarization technique (extractive vs. abstractive). Another key difference is that our information extraction methods only use textual and conversational features and are therefore applicable to conversations in different modalities (e.g., emails). In contrast, the system proposed in [7] is meeting specific, because it takes advantage of the audio-video recordings to better understand decision points.

The recent CALO meeting assistant [15] is also quite similar to ours. However, while they do perform an arguably more refined semantic analysis on the transcript, in term of topic identification and segmentation, question-answer pair

identification, as well as the detection of addressees, action items and decisions, they only provide extractive summaries of the conversation. More tellingly, their interface does not appear to be as visually sophisticated as ours, as it does not exploit the power on faceted search and tag clouds. One interesting feature of the CALO assistant is that it allows users to attach their own annotations to the transcripts, which is an interesting direction we could explore in our future prototypes.

5 Evaluation

Generally speaking, interfaces can be empirically tested with users in two ways [14]. In a comparison evaluation two interfaces are tested to see which one is superior with respect to a set of usability and performance metrics. In an assessment evaluation, only one interface is tested to verify properties of the interface, including: whether the user tasks are effectively supported, whether there are usability problems, and to identify relevant user differences in term of preferences and behaviors. In this paper, we present an assessment evaluation of our interface. In the user study, participants were asked to use the interface to browse and analyze a set of four long and complex human conversations (from the AMI corpus [4]). In particular, participants were asked to answer specific questions about the discussions that took place during the length of the conversations.

Thirty participants were recruited and compensated for the approximately 2 hours they spent on the study. A prize for the top three scorers was also offered to encourage people to get engaged in the task assigned. Two of the participants were excluded from the final analysis since they scored well below the average (more than two standard deviations). The remaining twenty-eight subjects ranged in age from 20 to 32. Twenty were male and eight were female.

Study Session	**Sample Tasks**	
Tutorial 1 conversation 2 participants 50 sentences	2. What kind of coat does Susan want to buy? Why did she not want a red coat or a tri-climate one? 3. What does Susan want for lunch? Did she consider other options? If so, what were the other choices?	
Practice 1 conversation 2 participants 100 sentences	1. What types of movies did Betty and Ronnie consider watching? What movie did they finally decide to watch? Why did they reject the other options? 2. What party are they planning to go to? Why is Betty worried about preparing for the party?	
Experiment		**Judges' Marking Scheme**
4 conversations 4 participants # of sentences min 363 max 1401 avg 901	2. A recent market survey revealed that battery life was an important feature considered by customers while buying a remote control. What are the options the design group considered about power of the remote? What was the final decision? Who proposed that solution?	Full marks: 3 4 options (0.5 marks each): a) double A, b) solar power, c) Lithium ion or long lasting or rechargeable and d) kinetic batteries; final choice kinetic battery (0.5 marks); proposed by the User Interface Expert (0.5 marks)
	3. One of the online customer reviews for the product read "A scroll-wheel like Apples ipod would have been cool for channel surfing!" Did the project group consider this option? If so, why was it discarded? What did they decide on in the end as the main way of interaction?	Full marks: 3 Yes, they discussed the scroll wheel (1 mark); decided to use push buttons instead (1 mark); rejected the scroll wheel because it is a) expensive and b) harder to control (1 mark if either reason mentioned)

Table 1. Study sessions, sample tasks and sample marking scheme for judges

We asked the participants to self-assess their comfort level using computers on a scale ranging 1 to 10 (where 1 meant they rarely used computers and 10 meant they could be considered expert computer users) and the median value was 8.5 out of a range of values from 3 to 10. All of the participants had normal or corrected vision.

The study procedure comprises three sessions - (a) a tutorial, (b) a practice and (c) an experiment session. The first two sessions were designed to ensure that the participant had acquired a common working understanding of the information displayed and of the interactive tools, before moving on to the experiment session. The instructions provided to the participants were carefully scripted for all three sessions to ensure that every participant was provided the same overview of the interface and how the different components can be used to derive the answers. At the beginning of the user study, we asked the users to fill up a pre-questionnaire to gather some background information.

In the tutorial session, the experimenter explained the different components of the interface to the user using simple tasks (see Table 1 for examples) on a short conversation between two participants. The user was advised to take some time to get comfortable using the interface at the end of this session. Next, in the practice session, the user worked on a slightly longer conversation of 100 sentences between two participants. The user was assigned two tasks (see Table 1) and was encouraged to work on her own. At the end of this session, the experimenter provided the user feedback on her performance for the assigned tasks.

The experiment session was timed with a limit of 1 hour. The user was assigned five tasks in a business oriented scenario (see Table 1 for two samples) based on a series of four related meeting conversations (ES2008 series of the AMI corpus) on the design of a new remote control by a group of four people participating with very specific roles. The four conversations were displayed on separate tabs, each tab containing all the basic controls and views as show in Figure 2. The settings and controls on one tab worked independently of the controls and settings on other tabs. The conversations ranged in length from a few hundred to more than a thousand sentences (see Table 1 for details). The five tasks in the experiment session were designed so that they could be answered independently of each other and in any order the user preferred.

At the end of the experiment we administered a questionnaire to get feedback on the usability of the interface, usefulness of its components and to get suggestions to improve the interface. The questionnaire consisted of Likert scale questions with scale value ranging from 1 to 5 and some open ended questions on different aspects of the interface.

All the sessions were carried out on a standard Windows 7 machine with 6 GB RAM, a 23 inch monitor and standard keyboard and mouse devices. During the experiment, interaction behaviors were automatically logged.

The answers of each participant were scored for accuracy by two human judges not associated with the project. Two sample marking schemes are shown in Table 1, besides the corresponding experimental tasks. The independent sets

of scores from the two judges were highly correlated (Pearson coefficient = 0.89). The average of the two scores was used in the final analysis.

6 Results

The aim of the user study was to assess the usability and effectiveness of our interface from different angles. Furthermore, since user accuracy on the task can be measured quantitatively, we are able to cluster users based on their performance and verify whether specific behavioral patterns and/or preferences for certain interface components are related to performing better or worse on the assigned tasks. For instance, we can answer questions like: Is it the case that the best performers used the Summary View more frequently? Or, is it true that the worst performers preferred more the Search Box. Answering these and similar questions on user behavior and preferences could inform the redesign of the interface and improve the training of new users.

6.1 Participants' Scores

As shown in Table 2, the performance of the participants varied substantially. We had three participants who achieved a perfect score of 12. To investigate user differences related to performance, we cluster our twenty eight participants into 3 mutually exclusive groups based on the scores they achieved: a high performer group containing nine people with performance score greater than 11, a mid-performer group containing ten people with performance score in the 8.5-11 interval, and a low performer group containing nine people with performance score within the 5.75-8.5 interval.

Min	**Max**	**Mean**	**Median**	**Std Dev**
5.75	12.00	8.87	9.50	2.83

Table 2. Summary statistics for scores of the 28 participants

6.2 Time Performance

Most participants took all the time allotted for the tasks. The ones who did not, finished only a few minutes earlier and did not specifically belong to any of the three groups based on the scores. Thus, we do not discuss time performance any longer.

6.3 Pre-Questionnaire

In the pre-questionnaire, we gathered information on the participants' gender, education level, comfort level in English, computer proficiency or familiarity. None of these properties showed a statistically significant correlation (at 0.05 significance level) with performance.

6.4 Post-Questionnaire

The questionnaire at the end of the experiment was designed to collect feedback on the overall usability of the interface, the utility of the different components and to gather suggestions on possible improvements. Table 5 shows the questions on the interface usability and perception of task context. Participants expressed their answers on a Likert Scale in which Strongly agree = 5; Strongly disagree = 1. In general, users found the interface intuitive, easy to use, and providing the necessary tools (median vales for Q1 and Q7 was 4). The participants also reported that they felt they were able to find relevant information quickly and efficiently (median value 4 for Q3), but sometime not all the needed information was available (median value 3 for Q2).

The fact that all the questions about the perceived task context had median value equal to 3 is an indication that the tasks were at an appropriate level of difficulty, not too easy, nor too difficult.

The post-questionnaire also gathered more specific feedback about the usefulness of the main components of the interface, and about the perceived accuracy of the information extracted by the NLP techniques. Results on these two aspects are shown in Table 3 and Table 4, respectively. From Table 3, it appears that, overall, a large majority of the participants found the three main components to be at least somewhat useful, with the Entity View being the most useful

Component	**Overall**			**High Perf.**			**Mid Perf.**			**Low Perf.**		
	+	+/-	-	+	+/-	-	+	+/-	-	+	+/-	-
Ontology View	9	14	5	2	5	2	3	5	2	4	4	1
Entity View	19	5	3	8	0	1	7	2	0	4	3	2
Summary View	14	10	3	6	3	0	2	5	3	6	2	0

Table 3. User feedback on the main components of the interface. Legend: + indicates Useful, +/- Somewhat Useful, - Not Useful. (Some overall totals do not sum up to 28 because of missing data from the questionnaire.)

Component	**Overall**			**High Perf.**			**Mid Perf.**			**Low Perf.**		
	+	+/-	-	+	+/-	-	+	+/-	-	+	+/-	-
Listed Entities	15	11	0	6	3	0	4	4	0	5	4	0
DA Tagging	13	6	8	4	3	2	5	3	2	4	0	4

Table 4. User feedback on the perceived accuracy of the two main Information Extraction tasks. Legend: + indicates Accurate, +/- Somewhat Accurate, - Inaccurate. (Some overall totals do not sum up to 28 because of missing data from the questionnaire.)

Questions	Mean	Median
Interface Usability		
Q1 I found the conversation browser intuitive and easy to use.	3.714	4
Q2 I was able to find all of the information I needed.	3.321	3
Q3 I was able to find the relevant information quickly and efficiently.	3.393	4
Q7 I had the necessary tools to complete the task efficiently.	3.571	4
Q8 I would have liked the conversation browser to have contained additional information about the conversations.	2.750	3
Q9 The interface quickly reflected the changes caused by interaction (changes caused when you select or unselect tags etc.)	4.036	4
Task Performance		
Q4 I feel that I completed the task in its entirety.	3.214	3
Q5 The task required a great deal of effort.	3.250	3
Q6 I felt I was working under pressure.	2.857	3

Table 5. Post-questionnaire questions. Assessments were expressed on a Likert Scale in which Strongly agree = 5; Strongly disagree = 1. For questions preceded by '*' lower values are better

one. Presumably, the Ontology View was rated by most user as somewhat useful or not useful, because as shown in Table 4 participants found the information displayed in the Ontology View (i.e., DA tagging) less accurate than the one in the Entity View (i.e., mid-frequency noun phrases). Based on this observation, future work should be focused on improving the accuracy of our classifiers.

With respect to differences among the three groups of participants based on performance, we did not notice any interesting difference or trend in either Table 3 or 4.

A final remarkable result from the post-questionnaire is that most participants did not find the abstractive summary very useful. Three participants explicitly mentioned that the abstractive summary is not well organized or useful, and needs more content. Others blamed lack of context for abstractive summaries being unsuitable as a browsing tool. In practice, only four participants tried out using the abstractive summary as a means of navigation and they clicked on less than 10 links each. This contrast with the results found in [11] in which abstractive summaries were assessed to be superior than extractive one. However, that study was quite different from ours. In [11], generic summaries were assessed based on their form and content, in our study they were assessed in the context of a realistic information seeking task.

6.5 Behavioral Patterns

The usage of different components of the interface was automatically logged based on clicking and scrolling behaviors. We were interested in answering the following key question: Do users, who perform better at the assigned tasks, follow different strategies (in term of components usage), than users who perform worse?

In general, our analysis of the interaction data indicates that high performers did not apply similar strategies. And the same is true for low performers. On the contrary, while some participants relied heavily on a single interaction techniques or a single view to explore the conversations, other adopted a more balanced strategy, independently from their performance level. For instance, even among the participants who achieve perfect score, one relied heavily on the generic Keyword Search (see participant P19 in Figure 3), while another one used the interface tools and views in similar proportions. For an additional example, look at Figure 4, which shows how two participants (P27 and P17), from two different performance groups relied heavily on the same interactive technique, namely the link between the extractive summary and the transcript for navigation.

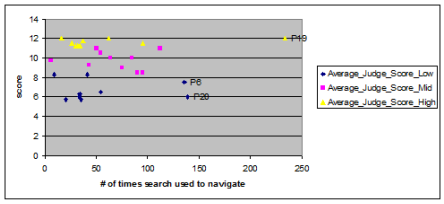

Fig. 3. Scatter plot for the number of times search button was used to navigate by a participant and participant's score. Participant 19 is one of the perfect scorers.

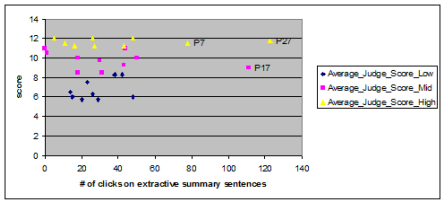

Fig. 4. Scatter plot for the number of time participants clicked on the extractive summary sentences and their score. P27, one of the high performers, and P17, one of the mid-level performers, both relied heavily on extractive summary for navigating within the transcript.

These findings appear to support the design choice of offering redundant functionalities within the interface. Different users seem to be able to achieve top performance in different ways, depending on their skills and preferences.

7 Conclusions and Future Work

This paper presents a visual, intelligent interface intended to help user analyze possibly long and complex conversations. The design of our interface aims to

effectively combine the power of human perceptual and cognitive skills with the ability of NLP techniques to mine and summarize text. Since we are targeting a large user population, with no, or minimal expertise in data analysis, we selected interface elements based on simple and common visual metaphors. Furthermore, to accommodate for user differences in such a large population, we provided users with the ability to satisfy the same information needs in different ways.

A comprehensive formative evaluation of the interface indicates that while we were quite successful in our endeavor, some problematic aspects still need to be addressed. On the positive side, not only users found the interface intuitive, easy to use and providing the necessary tools, but they also considered most of the key interface components to be useful. Furthermore, it seems that the choice of offering redundant functionalities was beneficial. The logged interaction behaviors reveal that different users actually selected very different strategies, even independently from their performance. And this was consistent with the high variability in the user preferences for the different interface components.

On the negative side, it seems that the NLP techniques generating the information displayed in the interface are still somehow unsatisfactory. More specifically, half of the users were at least partially bothered by inaccuracies in the classifiers that map utterances into the DAs they express. Furthermore, most users did not find abstractive summaries particularly useful. And abstractive summaries essentially represent the most sophisticated NLP component of our framework.

Based on these findings, our goal in the short term is to improve the NLP techniques. For the classifiers, higher accuracy could be achieved by applying more sophisticated machine learning techniques [5] and/or exploiting more informative sentence features (e.g., from a dependency parser). For the abstractive summarizer, improvements should be made to the various steps of the summarization process. As suggested by some users, the abstract summaries should be better organized and provide more specific information. For example, instead of generating sentences like: *The manager made negative comments about the display. She also made a decision on this matter.*, it should generate more informative sentences like: *The manager criticized the display layout and menus. They should be redesigned asap.*

Although users were in general quite satisfied with the Entity View, one possible improvement we plan to investigate is the presentation of the entities in coherent logical clusters; for instance, by following [9, 20]. Another possibility to provide a quick overview of the content of the conversation could be to apply more sophisticated topic modeling techniques. However, unfortunately, the accuracy of these techniques is still unsatisfactory [8].

While our approach relies on NLP techniques that can be applied to conversations in different modalities, including emails and blogs, the visualization of the conversation itself is so far limited to synchronous conversations with a linear thread, like meetings. In the future, we shall extend the Transcript View to display non-linear asynchronous conversation threads, typical of emails and blogs. For this, we will build on previous work, such as [16, 12].

References

1. Article on Tag Cloud, Wikipedia, http://en.wikipedia.org/wiki/Tag_Cloud
2. Bouamrane, M-M. Laz, S.: Navigating Multimodal Meeting Recordings with the Meeting Miner. In: Proceedings of FQAS, pp. 356–367. Milan, Italy(2006)
3. Albrecht-Buehlera, C., Watson, B., Shamma, D. A.: TextPool: Visualizing Live Text Streams. In: Proceedings of the IEEE Symposium on Information Visualization, pp. 215.1. Washington, D.C., USA (2004)
4. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI Meeting Corpus: A Pre-Announcement. In: Proceedings of MLMI, pp. 28–39, Edinburgh, UK (2005)
5. Criminisi, A., Shotton, J., Konukoglu, E.: Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. In: Foundations and Trends in Computer Graphics and Vision: Vol. 7: No 2-3 (2012), pp. 81–227
6. Dachselt, R., Frisch, M.: Mambo : A Facet-based Zoomable Music Browser. In: Proceedings of the 6th international conference on Mobile and Ubiquitous Multimedia, pp. 110–117. Oulu, Finland (2007)
7. Hsueh, P-Y., Moore, J. D.: Improving Meeting Summarization by Focusing on User Needs : A Task-Oriented Evaluation. In: Proceedings of the 14th international conference on Intelligent user interfaces, pp. 17–26. Sanibel Island, Florida, USA (2009)
8. Joty, S., Carenini, G., Murray, G., Ng, R.: Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 388–398. MIT, Massachusetts, USA (2010)
9. Kozareva, Z., Hovy, E.: A semi-supervised method to learn and construct taxonomies using the web. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1110–1118. Cambridge, Massachusetts (2010)
10. Murray, G., Carenini, G.: Interpretation and Transformation for Abstracting Conversations. In: North American ACL, Los Angeles, CA, USA (2010)
11. Murray, G., Carenini, G., Ng, R.: Generating Abstracts of Meeting Conversations: A User Study. In: Proceedings of the 6th International Natural Language Generation Conference, pp. 105–113. Trim, Co. Meath, Ireland (2010)
12. Pascual-Cid, V., Kaltenbrunner, A.: Exploring Asynchronous Online Discussion Through Hierarchical Visualization. In: Proceedings of 13th International Conference Information Visualisation, pp. 191–196. Barcelona, Spain (2009)
13. SIGIR'2006 Workshop on Faceted Search, https://sites.google.com/site/facetedsearch/
14. Stone, D., Jarrett, C., Woodroffe, M., Minocha, S.: User Interface Design and Evaluation (Interactive Technologies). Morgan Kaufmann (2005)
15. Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., Yang, F.: The CALO Meeting Assistant System. In: IEEE Transactions on Audio, Speech, and Language Processing, Volume: 18 Issue:6 (2010), pp. 1601–1611
16. Venolia, G. D., Neustaedter, C.: Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 361–368. Ft. Lauderdale, Florida, USA (2003)

17. Viegas, F. B., Wattenberg, M., Feinberg, J.: Participatory Visualization with Wordle. In: IEEE transactions on Visualization and Computer Graphics, Volume 15 Issue 6, pp. 1137–1144 (2009)
18. Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., Zhang, Q.: TIARA: a visual exploratory text analytic system. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 153–162. Washington, DC, USA (2010)
19. Wellner, P., Flynn, M., Guillemot, M.: Browsing Recorded Meetings with Ferret. In: Proceedings of MLMI, pp. 12–21. Martigny, Switzerland (2004)
20. Yang, H., Callan, J.: Ontology generation for large email collections. In: Proceedings of the 2008 international conference on Digital government research, pp. 254–261.
21. Yee, K-P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 401–408. Ft. Lauderdale, Florida, USA (2003)