A Computational Dialectics Approach to Meeting Dialogues Tracking and Understanding.

Vincenzo Pallotta
Faculty of Computer and Communication Sciences
Swiss Federal Institute of Technology, Lausanne, Switzerland
Vincenzo.Pallotta@epfl.ch

1 Introduction

Interaction through meetings is among the richest human communication activities. Recording meetings implies the storage and the structuring of a large set of heterogeneous information scattered over time and media. The raw data format from the various recording devices is not directly usable for the creation of indexes, nor for the content-based access to the relevant parts of the meeting recording.

The application scenario we envisage for meeting recording, understanding, storage and retrieval is the following: suppose someone has not attended a group meeting, but needs information about “what happened” at the meeting. In this situation the user might want to make queries about the meeting participants, about the issues that were discussed and the decisions that were made. Answers to these queries can be of different types. A list of participants can be an answer to the former question, whereas a written summary or an excerpt of the most relevant audio-video recording sequence can be the answer to the latter. The user might also be interested in accessing the documents related to the meeting, such as the agenda, reports, presentation handouts, etc.

How to retrieve the relevant information from the meeting recordings is itself an important issue in the design of the overall architecture for the application
scenario. On the one hand, and depending on the type of meetings, we need to specify the user requirements in terms of what the relevant information to be extracted from the meeting data is. On the other hand, the user might not be immediately aware of how the meeting recordings are structured, and he/she may be unable to formulate a direct query to the retrieval system. This last issue excludes the simple solution based on a search engine and forces us to consider a dialogue interface for query formulation and refinement. In the framework of IM2 project\(^1\) we also propose a web-based architecture for the multimodal access to the meeting recordings that allows us to structure the navigation through the meeting content as a dialogue and to combine several dialogue modalities, including voice.

In this paper we are mainly concerned with the higher-level annotations of meeting dialogues made on meeting's transcriptions. We assume for the time being the existence of a transcription, since we aim at designing tools for the extraction of semantic content from the linguistic surface expression. However, we are aware of the fact that additional information arises from several communication modalities, which will have to be integrated into the analysis tools we are currently designing. The overall goal of the project is to achieve a multimodal processing architecture where the linguistic dialogue analysis is one of the fundamental components.

### 1.1 The meeting data model

The data model required to store information about a meeting should take into account the following issues:

1. It must be instantiated by the information extracted from the analysis of the recorded meetings.
2. It must be indexed for the subsequent retrieval.
3. It contains temporal information.
4. It is described by concepts.

A possible approach is considering the meeting storage as a document base (a pool, or a repository), which contains sets of hypermedia documents. The nature of these documents is based on a formal ontology of entities. Due to

---

\(^1\) The National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management, in brief (IM)\(^2\), is aimed at the advancement of research, and the development of prototypes, in the field of man-machine interaction. For more information look at http://www.im2.ch
the dynamic nature of a meeting, we need to consider at least the following classes of entities:

- Individuals and objects
- Events
- Complex entities (aggregated structures).

The information contained in the meeting repository can be classified as static (a priori) and dynamic (on time).

Static information contains:

- The location of the meeting, the date, and the starting time.
- Information about participants (e.g. names, belonging institutions, roles, etc.)
- The layout of the meeting-room (e.g. participants seats, A/V recording channels, etc.).
- The information that is assumed of being initially shared by all participants (e.g. distributed documents, the meeting plan, the topics under discussion, etc.).
- The type of meeting (e.g. business, academic, political, etc.).
- The language spoken (e.g. native speaker, dialect)

Dynamic information contains time-stamped events:

- The participants' contributions (turns).
- Introduction of Referential Entities (i.e. entities evoked or referred to by the participants).
- The changes of topic under discussion (modelled, for instance, by a topic stack).

The information stored in the database can be organized at four different semantic levels:

1. Physical layer: the rough material annotated with absolute time and possibly with low-level extracted features without any semantic interpretation;
2. Factual layer: Information extracted from the rough material, e.g. from utterances, video-shots, prosodically separated speech units.
This information can be retrieved on the basis of the annotation of each unit;

3. **Thematic layer**: Information built by clustering together basic units of factual information into larger semantic units. For instance a set of dialogue turns may be structured into an episode which is a higher-level abstraction of dialogue;

4. **Rhetorical layer**: abstract relations between events or more complex event structures (e.g. episodes) of the meeting.

In modelling a multi-party dialogue for a meeting, one should be aware that there are two levels of task domain. One is the modelling of the topics the meeting is about, the other is the meeting itself. From the perspective of information retrieval from recorded meetings, the user might be interested in both aspects at the same time asking, for instance, “all the video sequences where person X where discussing with Y about the topic T”. Annotations must reflect at least these two dimensions, and then follow a suitable annotation schema.

On a first approximation, the meeting events can be viewed as relational database transactions of different types. For instance, it should be possible to consider all the contributions during a period in which a given topic has been under discussion just by selecting those contributions where their time-stamps are included in the interval of time in which the topic has been discussed. The meeting can be also viewed as a discussion forum where each participant contributes to the discussion by posting a message. Each posted message has an argumentative force with which the message is annotated. In this way, the meeting becomes a complexly structured (i.e. multi-threaded) document. By means of the identified argumentative force of each contribution, the message is appended to the document in the appropriate thread. The annotated meeting document can then be queried and navigated by the user in a usual way.

In multi-party dialogues such as meetings, the participants actually contribute to building up and updating the dialogue context. The dialogue context can be viewed as the “common ground” of the meeting in which all the participants share information. The dialogue context contains the dialogue state and can be described using a set of feature-attributes pairs. Since we are interested in the entire meeting, we would like to represent the temporal evolution of the context and to be able to take a “snapshot” of the context at a given time point.
The way the context is updated by dialogue contributions (a slightly more complex notion of dialogue act), can be represented by action occurrences (events). For each type of event, there are some update rules that can be applied. If no rule is specified, then the system simply records each contribution without attempting to reconstruct the dialogue context evolution. This means that in absence of update rules we are still able to extract some information about the meeting: the occurrence of the dialogue contributions (and their content).

The above perspectives can be integrated into a unique framework for the modelling of meeting situations. We achieve this goal by taking, as part of the dialogue context, the argumentative structure, which is incrementally built by assimilating knowledge extracted from dialogue acts, by means of update rules.

1.2 Meeting annotation

We might consider the analysis of meeting dialogues from different perspectives. On the one hand, we are interested in studying the occurring linguistic phenomena and in looking at correlations between the use of particular language patterns and a specific dialogue situation. On the other, we might look at the semantic structure of meeting dialogues, abstracting from their linguistic realization. If our goal is to exploit these types of analysis in our application scenario, we realize that the first type of analysis allows us to provide only a description of dialogues in terms of local intra-sentential phenomena like, for instance, the use of specific linguistic markers in the presence of given dialogue acts. The second type of analysis allows us to determine several abstract structures which are independent of their linguistic realization and which are aimed at providing a qualitative description of the phenomenon, as a whole. We intend to perform the latter kind of analysis and store the result in form of annotations.

Annotating a meeting entails the production of the content's meta-descriptions. Given a collection of time-stamped audio-video sequences, the transcription of the dialogues is of primary importance. It can be done manually by transcribers or automatically by a speech recognition system. The quality of the transcription, in terms of word-error rates, is a crucial issue on which all of the subsequent processing depends. It is apparent that a manual transcription provides us with reliable sources of information, while the performance of a speech recognition engine is very likely to be negatively influenced by the frequent noise situations that a meeting presuppose.
Furthermore, in order to capture the relevant features used to answer a given user query several additional layers of annotation are necessary. Depending on the type of the user query, we might need to extract features at different levels of abstraction and granularity; this point will be examined into details later.

The first type of annotation is based on the shallow dialogue model, proposed in (Armstrong et al., 2003). This model provides a simple logical structure for dialogues based on the following categories:

- a *dialog* is a non empty set of episodes; a new episode is identified by a topic shift.
- an *episode* is a non empty set of turns; a new turn is introduced by a speaker change.
- a *turn* is a non empty sequence of *utterances*.

Topic shift detection techniques can be used in order to find the boundaries of topically homogeneous segments in texts/speech transcriptions. A problem in natural multi-party dialogue is that the topic shifts are not so abrupt as opposed to picking data from different documents. Each turn is annotated by one or more dialogue acts highlighting the communicative function of an utterance. The set of dialogue act labels is based on Switchboard/DAMSL guidelines (Core and Allen, 1997) and it is currently used for the annotation of the ICSI corpus of meeting dialogues that we use for our tests. DAMSL dialogue acts can be roughly classified into three main categories:

1. *Communicative-Status*: Uninterpretable, Non-verbal, Self-talk, etc.
2. *Forward-Communicative-Function*: Statement, Influencing-addressee-future-action (e.g., questions), Committing-speaker-future-action (e.g. offers), Other-FCF (e.g., apology, thanks)
3. *Backwards-Communicative-Function*: Agreement, Understanding, etc.

Extensions of this set towards multimodal dialogue acts are currently under investigation by members of the IM2 project. Aggregations of simple dialogue acts are provided by the *adjacency pairs* proposed for conversational analysis by (Sacks et al., 1977) in order to define the functional link between two utterances. Examples of adjacency pairs are:
• invitation <-> accept / decline
• question <-> answer (comply, supply, evade,...)
• request <-> accept / turn down, delay, promise, ...
• greeting <-> greeting

Adjacency pairs will allow us to answer user’s questions about the unstructured knowledge on the meeting events such as, for instance:

• Show me all sequences where B refused to do what A requested him to do.
• What questions by X were still left unanswered after this meeting?
• Let me see the moment when X and Y strongly disagreed.
• On which points were there no objections?

In addition to the shallow model, which will be mostly automatically extracted, we also consider the adoption of a deeper representation. It will be produced with more manual intervention, and will provide the system with more possibilities for the users to identify the desired segments of the meeting. In fact, the main limitation of the shallow dialogue model is that a single utterance may have multiple communicative functions and that there is no trace of participants’ intentions. In multimodal dialogues, for instance, there are other types of communicative actions besides utterances, e.g. agreement by applause, disagreement by gesture or facial expressions. Moreover, the model does not take into account the formation of opinions by hearers about speakers, and, more important, it does not highlight the social behaviour of the participants, nor their role in the deliberation process.

In order to overcome the above limitation, we propose to consider meeting dialogues from the Collaborative Decision Making (CDM) perspective. A meeting can be viewed as a multi-party (multi-agent) decision making process: a collaborative process, where agents follow a series of communicative actions in order to establish a common ground on the dimension of the problem. The main dimensions of CDM process are:

• An overall task goal;
• A set of alternatives;
• A collection of choice criteria (perspectives and preferences) settled upon the participants;
• A decision (or evaluation) function that combines criteria to judge the alternatives.

<table>
<thead>
<tr>
<th>Type of Dialogue</th>
<th>Initial Situation</th>
<th>Participant Goal</th>
<th>Goal of Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persuasion</td>
<td>Conflict of opinions</td>
<td>Persuade other party</td>
<td>Resolve or clarify issue</td>
</tr>
<tr>
<td>Inquiry</td>
<td>Need to have proof</td>
<td>Find and verify evidence</td>
<td>Prove (disprove) hypothesis</td>
</tr>
<tr>
<td>Negotiation</td>
<td>Conflict of interests</td>
<td>Get what you most want</td>
<td>Reasonable settlement</td>
</tr>
<tr>
<td>Information seeking</td>
<td>One party lacks information</td>
<td>Acquire or give information</td>
<td>Exchange information</td>
</tr>
<tr>
<td>Deliberation</td>
<td>Dilemma or practical choice</td>
<td>Co-ordinate goals or action</td>
<td>Decide best course of action</td>
</tr>
<tr>
<td>Eristic</td>
<td>Personal conflict</td>
<td>Verbally hit out at opponent</td>
<td>Reveal deeper basis of conflict</td>
</tr>
</tbody>
</table>

Table 1. The Walton and Krabbe's classification of dialogue situations.

The above perspective assumes that meetings are rational human activities as pointed out in the Ralph Johnson’s Manifest of Rationality (REF):

"...[The meeting] is a socio-cultural activity of constructing, presenting, interpreting, criticizing, and revising arguments for the purpose of reaching a shared rationally supported position on some issue”.

This definition focuses on the processes which take place during meetings and how these processes contribute to the accomplishment of a joint goal.

According to a classification provided by NIST\(^2\), these three types of meeting scenarios might contain the following CDM processes:

\(^2\) Meeting scenarios from the National Institute of Standards and Technology: http://www.nist.gov/speech/test_beds/mr_proj/data_collection/scenarios.html
1. Staff Meetings: Participants discuss real technical issues, brainstorm ideas and make decisions. They include also planning, negotiation and brainstorming.

2. Information exchange and decision-making meetings. For instance, office furnishing: An expert will help participants to choose office furniture, carpet, etc. for an office. They also include brainstorming.

3. Information gathering and decision-making meetings. For instance, shopping on-line where participants search the Web and collaborate to purchase a digital camera. This type of meeting includes also negotiation.

Depending on the type of meeting it is useful to consider a taxonomy of dialogue situations such as the one of (Walton and Krabbe, 1995), shown in Table 1. This classification is particularly useful to formally define the protocols of meetings where a decision-making process is taking place. The main rationale of having a formal description of such decision processes is to provide answers to complex questions from the user. These questions are mainly those pertaining to the outcome of the discussion in terms of the arguments invoked, questions raised, and consensus achieved on the discussed issues. There are some examples of possible questions types:

- How did we arrive at the conclusion that X?
- What were the objection against the proposal that Z?
- What was the position of X on subject Z?
- For which open question there was no solution adopted? Why? What are the open questions for a next meeting?
- Which criteria were chosen to take the decision D1?
- Which criteria were invoked by the members who disagreed on the decision D1?
- Give me for each topic the list of people that said something related to it. Then give me their main point (agreement/disagreement with X).
- When did X contradict himself about the issue I.

2 The construction of a meeting dialogue ontology

Our research plan within the IM2 project focuses (among other issues) on the automated tracking, topical indexing, and summarization of multimodal
human-to-human interactions within the application scenario outlined in the previous section. This research requires the definition of precise annotation guidelines, the transcription and annotation of a large number of recorded meetings, including their semantic/conceptual annotation and the annotation of the dialogue structure. In a first phase, a large portion of the resources required for the software prototypes will be handcrafted, relying on available test data. However, it is our goal to automate the annotation process, and to provide a large-scale ontology for this process. It is therefore of central concern to progressively set up a framework within which the production of the ontology is carried out in a systematic and at least semi-automated way. A dialogue-type-specific ontology can be used to represent and store the information related to the purpose and nature of the dialogue, regardless of the technical domain in which the dialogue takes place. This type of ontology expresses conceptual information about the roles of participants and the type of dialogue acts that take place between them.

Any attempt to access and process the informational content of the meeting benefits from the mapping between that particular content and the corresponding specialized ontology. For instance, intelligent access to the dialogues from a business meeting implies that these dialogs can be first processed - e.g., stored, indexed, summarized - according to domain-specific and dialog-type specific ontology. Therefore, this ontology must be somehow formalized and built in a flexible and efficient manner.

We compare our efforts to that of the MPEG7 consortium\(^3\) (REF), aimed at creating a standard for the description of semantic content of audio-visual documents. MPEG-7 is formally called "Multimedia Content Description Interface", and describes means of attaching metadata to multimedia content. MPEG-7 is a standardized description of various types of multimedia information. This description is associated with the content itself, to allow fast and efficient searching for material that is of interest for the user. MPEG-7 specifies a standard set of descriptors that can be used to describe various types of multimedia information. MPEG-7 also standardise ways to define other descriptors as well as structures, Description Schemes (DSs) for the descriptors and their relationships. These elements describe the Structure (e.g. regions, video frames, and audio segments) and the Semantics (e.g. objects, events, abstract notions). The functionality of each of these classes of DSs considers structural aspects, describing the audio-visual content from the


10
viewpoint of its structure, and the conceptual aspects, describing the audio-visual content from the viewpoint of real-world semantics and conceptual notions.

3 Argumentative structure of meetings

In order to answer the types of questions exemplified in the previous subsection, we need to further annotate meeting recordings with appropriate meta-description. In other words, we need to annotate the parts of the meeting where the decision process takes place with a suitable "argumentative structure". This information will serve as an additional index for retrieving the relevant information about the discussion outcomes. Moreover, the argumentation structure may be used as the underlying template to building meeting summaries.

A proposal for an Argumentation Mark-Up language has been proposed in (Delannoy, 1999) providing a set of XML tags to pre-process text containing arguments in order to build summaries. We believe that this model is not sufficient for the meeting dialogues, since it only highlights argumentative rhetorical relations in monologues.

A simple model, showed in figure 1, of an argumentative structure is the "Issue Based Information Systems" (IBIS), proposed by (Kunz and Rittel, 1970) and adopted as a foundational theory in some computer-supported collaborative argumentation (CSCA) systems such as Zeno (Gordon and Karacapilidis, 1999), HERMES (Karacapilidis and Papadias, 2001), Questmap (Conklin et al., 2001), and Compendium (Selvin, 2001). The main goal of IBIS is to reach a rational agreement on topic issues by following a pre-defined argumentation protocol. We adopt this model as a first basic
model for the description of the argumentative structure of decision meetings. The model captures and highlights the essential lines of a discussion in terms of what issues have been discussed and what alternatives have been proposed and accepted by the participants. The discussion is thus represented by a discussion graph in which each node is linked to the related segments in the meeting recordings. In order to automatically generate this structure we need to extract argumentative acts from the transcription. Argumentative acts are interpreted as update actions to the current argumentative structure.

The development of the argumentation structure is a dynamic process which itself needs to be modelled. In (Gordon and Karacapilidis, 1999), the process of proposing and arguing over alternatives is modelled by a state transition graph. This simple model states a strict protocol, which constrains the interaction between participants, but hardly scales up from computer-mediated discussion to real life unconstrained meetings. We believe that a more general model of argumentation dynamics is required in order to be capable of dealing with multi-party and face-to-face deliberation dialogues.

An important step towards this direction is represented by the work of Hitchcock et al. They present in (Hitchcock et al., 2001) a framework for deliberation dialogues, which generalize to the underlying dynamic model of dialogue games (Amgoud et al., 2000), initially conceived to model persuasion, inquiry and information seeking dialogues. This framework includes a set of dialogue acts, which can be interpreted as transitions from one stage to another in the dialogue game. Unfortunately, the dialogue moves are not interpreted as constructors of an argumentation structure; they are rather used to update the participants' commitment store. The agreement on some issue is reached when all the agents are committed on the same alternative.

In (Ballim and Karacapilidis, 1998), a set of dialogue acts involved in the argumentation process, that is, argumentative acts, are used to analyze deliberation dialogues by tracking the participants' mental states. The dialogue acts are described as plan operators over mental states. The main goal of this kind of analysis is to improve the quality of support to decision making in CSCA by detecting contradictions, and coherence in the dialogue (e.g. if a participant believes P, and believes that Q refutes P, then he/she cannot reasonably propose Q without first discarding P). From the perspective of meeting recording, this kind of analysis might be useful to provide information about the participants' goals and intentions (not explicitly stated) at the beginning of a dispute, and possibly answer related user's questions. This can be obtained by applying plan-recognition techniques in order to infer
reasonable mental conditions on the meeting participants. Argumentative acts have effects in the participants' mental state and update the common ground (mutual beliefs, joint goals and intentions); some argumentative acts create expectations (e.g. askinfo, askopinion, askact, acknowledge). Others make explicit the speaker mental attitudes (e.g. inform, consider, clarify, agree, disagree), create/destroy hypothetical issues (e.g. consider, discard), or generate a negotiation (e.g. propose, corroborate, challenge, counter-offer, compare).

One main difference between CSCA and meeting recordings is that, in the former case, the argumentative structure is used to constrain the interaction whereas, in the latter, we aim at deriving the corresponding argumentative structure (if any) by observing a recorded interaction. It is apparent that the second problem is harder since one has to infer the causal relationships of dialogue events without knowing the participants' intentions and goals. Modelling the dynamics of argumentation also means dealing with multimodal knowledge about the dialogue events, since some argumentation acts can be stated in other modalities (e.g. agreement by silence or by applause, disagreement by laughing).

The importance of tracking collaborative argumentation of discussion meetings has a central importance for building what Duska Rosemberg calls "Project Memories" in (Rosemberg and Sillince, 1999). The construction of projects memories is similar to the annotation of meetings by their argumentation structure since it highlights not only "strictly factual, technical information", but also relevant information about the decision making process.

Discussion meetings are likely to include negotiation and agreement steps. These are the main components of an argumentation structure and they can be annotated without having to build an entire argumentation structure. This approach has been pursued for the annotation of a corpus of computer-mediated problem-solving dialogues. Although the dialogue setting is rather constrained, it shows important issues, which can be relevant in building an argumentation structure of meeting dialogues. The model proposed in (Di Eugenio et al., 2000) for modelling the COCONUT task is obtained by the integration of the IRMA agent architecture (Pollack, 1992) and the Clark's theory of common ground (Clark, 1996). The former is used as a computational model for the rational agent decision making process, while the latter models the use of language for dialogue grounding. The adoption of the Clark's Common Ground theory is motivated by the need of modelling the
agreement process by the notion of joint commitment rather than that of proposal-acceptance. This hypothesis is validated by a correlation study on the COCONUT corpus, which has been annotated by an extended DAMSL tag-set, which includes categories of the IRMA model.

The approach highlighted here is directly in line with the ongoing research within the Semantic Web\(^4\) framework. More specifically, this research addresses the need to establish links and to integrate information coming from ontology elaboration on the one hand, and more classical lexical knowledge base development on the other. The argumentation schema we propose as the starting point for the construction of a meeting ontology is based on RDF\(^5\) schema and reflects the substantial aspects of the IBIS model.

4 Conclusions

The automatic construction of argumentative structures from meetings is a long term goal, while manual construction seems to be more reasonable in the short-term. Nevertheless, there are some important steps, which can be performed automatically. Recent works (McCowan et al., 2002) made by partners in the IM2 project have been focused on stochastic models for the automatic detection of meeting episodes based on the combined extraction of multiple audio-visual features from meeting recordings. Although the studies are preliminary and accuracy result not yet satisfactory, they recognize the importance of detecting discussion situations. This suggests that not all the types of meetings will require an argumentation analysis and not every part of the meeting will contain arguments. Indeed, the automatic detection of candidate segments where an argumentation is likely to be found could be of great help to the human annotator. Moreover, in an adaptive annotation tool (Ballim et al., 2000), topic segmentation can be used to propose annotation of issues, and the presence of a high number of propose-accept or propose-reject adjacency pairs may signal a segment which contains an argumentation act.

Another important advantage of having an additional layer of annotation is the improvement of precision during retrieval. We believe that it will beneficial to the end user to include the argumentative structure in the dialogue model providing an additional conceptual layer for semantic content representation. Both queries and results would better be formulated and presented if the system can rely on a well-defined structure of meeting events.

\(^4\) http://www.semanticweb.org
\(^5\) http://www.w3.org/2001/sw/RDFCore/
5 Acknowledgements

This work is part of a Swiss National Science Foundation project, (IM)2, i.e. Interactive Multimodal Information Management (http://www.im2.ch). More specifically, the present work belongs to the Multimodal Dialogue Management (MDM) module (http://www.issco.unige.ch/projects/im2/mdm).

I would like to thank all the members of the MDM group for their contribution to the development of the ideas presented in this article. I am also grateful to the ICSI Meeting Recorder project for access to their data. Thanks also to Violeta Seretan and Giovanni Coray for their comments and suggestions.

References


