User Query Analysis for the Specification and Evaluation of a Dialogue Processing and Retrieval System

LISOWSKA, Agnès, POPESCU-BELIS, Andréi, ARMSTRONG, Susan

Abstract

This article describes an experiment in user query elicitation for the design of a multimodal meeting processing and retrieval system (MPR). In the experiment, participants are asked to choose between several scenarios of use of an MPR system, then formulate (on paper) queries to the system within the context of their chosen scenario. The analysis of the queries provides us with an initial set of requirements for the design of an MPR system, which will be used to confirm a priori design considerations, and suggest improvements to existing interfaces. This elicitation-design-evaluation process will be iterated, where the next phase will involve experiments using the Wizard-of-Oz methodology.

User Query Analysis for the Specification and Evaluation of a Dialogue Processing and Retrieval System

Agnes Lisowska, Andrei Popescu-Belis, Susan Armstrong
ISSCO/TIM/ETI, University of Geneva
40 bd. du Pont d'Arve, CH-1211, Geneva, Switzerland
{Agnes.Lisowska, Andrei.Popescu-Belis, Susan.Armstrong}@issco.unige.ch

Abstract

This article describes an experiment in user query elicitation for the design of a multimodal meeting processing and retrieval system (MPR). In the experiment, participants are asked to choose between several scenarios of use of an MPR system, then formulate (on paper) queries to the system within the context of their chosen scenario. The analysis of the queries provides us with an initial set of requirements for the design of an MPR system, which will be used to confirm *a priori* design considerations, and suggest improvements to existing interfaces. This elicitation-design-evaluation process will be iterated, where the next phase will involve experiments using the Wizard-of-Oz methodology.

1. Introduction

This paper reports on work in multimodal dialogue management (MDM)[1], within the Interactive Multimodal Information Management (IM2)[2] project, in the domain of recorded meetings. One of the long term goals of this work is the development of a multimodal meeting processing and retrieval (MPR) system. It is important to note that multimodality in such a system plays a role on two levels: in the content of the meetings themselves since human-human interaction is by nature multimodal, and in the way in which a person can access that content via a multimodal interface to the system. Our work takes both of these levels into consideration.

Since the resulting application should satisfy realistic user needs, it is essential to ground its design in real user requirements for the intended task. In this article we describe an experiment in which queries were elicited from potential users of an MPR application, then proceed to analyze the constraints they impose on the design and evaluation of the system.

2. Software Lifecycle: from User Modelling to Evaluation

Our approach to user modelling and evaluation is based on standard HCI practices (Dix et al., 1998) and on the ISO/IEC 9126 and 14598 sets of standards. With respect to the role of quality in the software lifecycle (ISO/IEC 14598-1, p.12), the formal design of an application should start with a precise empirical assessment of the user requirements for it. From these requirements, a set of quality requirements can be defined for the application (i.e., the behavioural features of the system), which in turn generate an internal set of specifications, on which a first version of the system can be constructed. Evaluation should take place at each of these three levels: (a) internally, by evaluating internal parameters of the system such as language models, rule patterns, etc; (b) externally, by running the system on sample data and measuring the level of performance; and (c) in use, by observing the satisfaction of real users interacting with the system. The specification-design-evaluation cycle can be iterated to refine the resulting software.

3. Collection of Requirements

In order to gain insight into what aspects of meetings users of an MPR system might want to know about, an initial informal study was performed where sets of possible queries to an MPR system were elicited from participants in the IM2 project. The subjects were simply told to formulate queries that would enable them to find out "what happened at a meeting" that they did not attend, or alternatively, review specific points of a meeting they attended. Although about 200 queries were produced in this way, it was felt that the possibility of the subject group introducing bias to the data set and the relative openness of the task being performed merited a refinement of the parameters of the study.

Consequently, a second more principled study was conducted (Lisowska, 2003) with a wider range of participants in order to control for bias, and a more constrained set of instructions to allow for a more coherent data set, facilitating analysis. It is the results of this second study that are described in this paper.

The study was done using a questionnaire, distributed by e-mail to the participants, which first outlined the IM2 project, its aims, and the intended system, then went on to describe four scenarios in which someone might be using the system. These scenarios were:

- an employee who has missed a meeting on a project they are involved in and wants to catch up (12 participants)
- a new employee who is using the system to familiarize themselves with a project that they will be involved in (7 participants)
- a manager who is tracking the progress of a project (4 participants)
- a manager who is tracking employee performance (5 participants).

The participants were asked to choose one of these scenarios, place themselves in the role described in it, and write down queries, just as they would if they were actually interacting with the system. The final part of the questionnaire asked some general questions about the

[1] http://issco-www.unige.ch/projects/im2/mdm/
[2] http://www.im2.ch

participants such as their profession, computer experience, and native language. About 300 queries were gathered from participants both involved and uninvolved in the IM2 project (14 participants from each category).

4. Analysis of User Queries

Our first goal was to define a number of classes and to sort the queries according to the requirements they induce. The various requirements that can be inferred concern:

- data recording (especially regarding different modalities);
- processing of the recordings (particularly of dialogue);
- linguistic processing of the queries themselves;
- domain specific support tools (such as ontologies).

There are two broad classes pertaining to meetings and meeting dialogues into which a vast majority of the queries can be placed:

(a) Elements related to the interaction among participants:

- acceptance/ rejection, agreement/ disagreement
- proposals, argumentation (for and against)
- assertions, statements
- decisions
- discussions, debates
- reactions
- questions
- solutions

(b) Concepts from the meeting domain

- dates, times
- documents
- meeting index: current, previous, sets
- participants
- presentations, talks
- projects
- tasks, responsibilities
- topics

Moreover, a single query can sometimes simultaneously belong to several groups from either class. The initial categorization, still subject to discussion, does shed light on four important points. The first is that the two classes, (a) and (b), are probably not completely disjoint. For example queries about the assignment of responsibility during a meeting belong to (a) while queries about pre-established tasks belong to (b). The second point is that the number of queries per class differs, as overall more queries belong to class (b) than (a). The third point is that a surprisingly high number of queries require little processing of meeting data. Finally, some queries refer to "absent" elements, that is, items from (a) or (b) that are not present in a meeting (but could or should have been) - for example, "*Which point of the agenda wasn't discussed due to lack of time*?" The ramifications of these points need to be investigated in more detail.

5. Implications of the Analysis

The analysis of the elicited queries points to a number of system requirements, both from the meeting recording and processing, and from the retrieval perspectives. While not all of the requirements are feasible at present, the selection below is based on feasibility and potential for research.

5.1. For Meeting Recording

While we believe that the nature of the questionnaire format is not conducive to determining precise requirements concerning media formats, the initial indications that follow are likely to hold regardless of the elicitation means. Both audio and video footage of the meetings are required, and potentially, both these media will need to be enriched by annotations in order to make them accessible for retrieval. These annotations may be directly on the media, or as in the case of audio, on transcriptions. Additionally, documents that accompany the meeting or are mentioned in it (such as presentation slides, agenda, reports, notes taken during the meeting, whiteboard diagrams, etc.) were often the subject of queries and would have to be included in the data. Semantic analysis of those queries also indicated that the links relating those documents to the portion of the meeting in which they appear and to their owner are also necessary.

5.2. For Meeting Processing - Relevant Dialogue Phenomena

The analysis of the queries confirmed that meetings stored in the system need to be processed in order to make their content accessible for retrieval. One particularly necessary aspect of this processing is at the dialogue level, both in terms of content and structure.

5.2.1. Shallow Dialogue Analysis

The queries highlight the importance of dialogue elements that could be extracted using shallow dialogue analysis. One of the most frequent elements is topics, which requires the segmentation of a dialogue into topic-coherent episodes, and the assignment of titles or keywords to them. Also, references to named entities such as times/dates and people are required to answer a significant proportion of the queries, as are references to documents.

5.2.2. Dialogue Structure

Theories of dialogue do not offer a complete framework for dialogue understanding, but rather focus on partial aspects such as dialogue acts or turn-taking. Therefore, to fulfil the observed user requirements, coherent sets of features should be considered, which should be based on existing theories, feasibility, and empirical evidence.

Queries that require an understanding of dialogue structure were somewhat less frequent than expected, but still comprised a sizeable proportion of the queries. Several dialogue acts (Popescu-Belis, 2004) are explicitly requested in the queries, such as

- statements (24 times)
  e.g. "*Has somebody ever said when the project was due to end?"*
- proposals (12 times)
  e.g. "*Which proposal was accepted without any discussion?"*
- questions (5 times)
  e.g. "*Did somebody answer the question I asked last week?*"

Other queries require the detection of adjacency pairs. Elements such as acceptance/rejection or reactions play this role only as part of a higher level structure, for example a question/answer adjacency pair. Less tractable

are queries about discussions, argumentation and decisions, which seem to require a high level understanding of the dialogue. These types of queries are nevertheless quite frequent, which shows that robust methods should be developed for detecting them in meetings.

5.3. For Retrieval – Query Processing

Analysis of the queries at the semantic level indicates that in order to find the parts of a meeting that the queries refer to, both query processing and domain specific support tools are needed.

5.3.1 Linguistic Processing of the Queries

The MPR system we envision allows the user to perform directed searches that we believe will be language based since language is a natural interactive modality for human beings. However, in order for the system to respond in an appropriate manner, the queries themselves need to be processed. This can be performed using standard natural language processing tools adapted to the meeting domain.

5.3.2 Domain Specific Support Tools

The queries pointed to the need for two such tools. The first of these is a domain-specific ontology - an ontology of meetings. While ontologies are notoriously hard to construct, we believe that the query set obtained in this study can be used as a seed ontology that could subsequently be augmented and expanded as further experiments are run.

Another related yet simpler tool that is required is a mechanism for resolving temporal references in queries such as '*When is the next meeting?*', where the meeting that is referred to needs to be determined based on a derivable date and the definition of the word 'next' in the context of meetings.

5.4 Indications for Overall System Design

The multi-dimensional analysis of the queries highlights the high level of interdependency between all of the components and aspects of the system. This interdependency implies that the components cannot be developed on a stand-alone basis if the system as a whole is to meet overall user requirements, but rather need to be developed under careful consideration of the other components with which they interact, which introduces additional complexity into the design and software architecture of the system. For example, the queries that users pose determine to a large extent the types of annotations that are necessary on the data, while the annotations in their turn drive the structure of the database and how it can be accessed.

6. The Role of Query Elicitation in Subsequent Evaluation

In accepted evaluation practice, data that is used for the specification of system requirements should not be used for testing the implementation of that specification. However, in the particular case of the MPR system and the set of elicited queries, we believe that the data that is used for the specification of the system can in fact be used to a certain extent for evaluation purposes as well. The elicited queries serve as a benchmark. They are a projection of all of the desired requirements, and do not take into consideration the limitations imposed by the state of the art of the technologies involved in building the system. We will call the set of these queries the *superset*.

When a system is actually being developed, the technological constraints (for example from HCI, NLP or database technologies) have to be taken into consideration, and given the current state of the art in various related areas, only a *subset* of the superset can realistically be used for MPR system specification.

The superset can then be used to test the breadth of the coverage of the system and to evaluate progress that is being made, particularly when technologies are added to the system as they become available. In the latter case, it is unlikely that an existing MPR system will be redesigned from scratch taking into account the new technology, in which case the superset provides a convenient measure of comparison between the current system and its predecessor.

7. Impact on Current and Future Work

The results of the query elicitation study show that the static nature of the questionnaire format is insufficient to elucidate all the types of requirements needed for the system design. For example, the query set gave no indications as to the modalities that users prefer to use to access the meetings and formulate their queries, nor the modalities in which they prefer to receive the data that serves as a response to their queries.

There were two additional factors that pose doubts as to the sufficiency of coverage of the query set. The first is the limited size of the data set. The second is that the description of the imagined system, which had to be relatively detailed in the questionnaire in order to properly set out the task, could have biased the content of the queries expressed. We believe that the remedy to all of these shortcomings can be found in the next iteration of the design process, described below.

7.1 Current Interfaces

Two interfaces for dialogue retrieval have already been developed by the MDM group in the context of the IM2 project (Armstrong et al., 2003). The first is a direct form-based interface to a database of processed meeting dialogues, and allows retrieval of utterances and their context based on a conjunction of utterance parameters ("selection") using the traditional keyboard and mouse modalities. The second interface, built on the same data structure as the first, has improved interaction capabilities, in particular multimodal input (speech, text, and mouse). Both interfaces give access to the elementary shallow dialogue structure of the meetings. The design and implementation of these interfaces preceded the study described in this paper, and is primarily data driven rather than user driven, with user requirements being elucidated at the intuitive rather than empirical level.

7.2 Reorienting the Current Interfaces

Now that a principled preliminary study of user requirements for this particular domain has been done, we are in a position to reorient the existing interfaces, which already reflect current underlying system constraints, to accommodate real user needs. In doing so, we will be

taking the first step in actively fusing user requirements and system constraints.

Once the interfaces are implemented, another iteration of the design cycle can begin. In this iteration, the interactive nature of the existing interfaces can be exploited to gain insight into the different input and output modalities that users prefer during interaction, to further refine the interfaces and means of interaction themselves by examining both directly in the context of use, and finally, to evaluate the breadth of the coverage of the query sets by continuously building on and analysing them as new queries are put to the system and logged during the experiments.

7.3 Wizard of Oz experiments

There are two fundamental problems with designing and testing highly multimodal systems such as an MPR. The first comes from the user requirements perspective and involves determining which user requirements are the most relevant and technologically viable. Moreover, it is desirable to avoid implementing designs that are theoretically interesting but are revealed to be unsuitable in a real use context. The second comes from the technological constraints perspective. One wants to avoid putting a priori limitations on the means of interaction that are available in an interface as they risk constraining both the task and interactions in such a way that the interaction becomes unnatural and as a result, the data gathered is unrepresentative of real use.

In order to avoid these pitfalls, we propose to run Wizard-of-Oz experiments, where functionalities that have not yet been implemented are provided by a hidden human controller (the wizard) in such a way that the user believes they are interacting with an autonomous system (Dahlbäck, Jönsson, & Ahrenberg, 1993; Salber & Coutaz, 1993). Such studies allow greater flexibility in gathering a wider variety of user requirements, particularly of the multimodal kind, while at the same time reducing the risk of running into the problems described. Furthermore, they allow for both elicitation and analysis of phenomena such as chained queries (i.e. queries that refer back to previous queries or results) which were only occasional in the questionnaire context but are expected to be much more common in an interactive context. These phenomena, and the way in which they are accounted for by the system (for example by the dialogue manager) could have a significant influence on the design of the interface itself.

8. Acknowledgements

The work presented here is part of the Swiss NCCR on "Interactive Multimodal Information Management" (IM2, http://www.im2.ch), funded by the Swiss National Science Foundation. The work pertains specifically to the IM2.MDM module, "Multimodal Dialog Management" (http://www.issco.unige.ch/projects/im2/mdm).

9. References

Armstrong, S., Clark, A., Coray, G., Georgescul, M., Pallotta, V., Popescu-Belis, A., Portabella, D., Rajman, M. and Starlander, M. (2003) Natural Language Queries on Natural Language Data: a Database of Meeting Dialogues. Proceedings of NLDB'2003 (8th International Conference on Applications of Natural Language to Information Systems), Burg (Spreewald), Germany, p.14-27.

Dahlbäck, N., Jönsson, A. and Ahrenberg L. (1993). "Wizard Studies – Why and How". In W.D. Gray, W.E. Helfley and D. Murray (eds), Proceedings of the Workshop on Intelligent User Interfaces, pp. 193-200.

Dix A., Finlay J., Abowd, G., Beale, R. (1998). "Human Computer Interaction". Prentice Hall, Harlow, U.K.

ISO/IEC (2000). ISO/IEC 14598-1: Information Technology – Software Product Evaluation – Part 1: General Overview. International Organization for Standardization, Geneva.

ISO/IEC (2001). ISO/IEC 9126-1: Software Engineering – Product Quality – Part 1: Quality Model. International Organization for Standardization, Geneva.

Lisowska, A. (2003). Multimodal Interface Design for the Multimodal Meeting Domain: Preliminary Indications from a Query Analysis Study. IM2.MDM Internal Report IM2.MDM-11, November 2003. http://issco-www.unige.ch/projects/im2/mdm/docs/MDMReport-11-AL.pdf

Popescu-Belis, A. (2004) Abstracting a Dialog Act Tagset for Meeting Processing. Proceedings of LREC 2004, Lisbon, Portugal.

Salber, D., and Coutaz, J. (1993). "Applying the Wizard of Oz technique to the study of Multimodal Systems". Proc. of EWHCI'93 (3rd International Conference East/West Human Computer Interaction), Moscow. L. Bass, J. Gornostaev, C. Unger, eds. Springer Verlag Publ. Lecture Notes in Computer Science, vol. 73. pp. 219-230.