Task-based evaluation of meeting browsers: from task elicitation to user behavior analysis

Andrei Popescu-Belis[1], **Mike Flynn**[1], **Pierre Wellner**[1], **Philippe Baudrion**[2]

[1] IDIAP Research Institute
Av. des Prés-Beudin 20
CH-1920 Martigny

[2] ISSCO, University of Geneva
Bd du Pont-d'Arve 40
CH-1211 Geneva 4

andrei.popescu-belis@idiap.ch, mike.flynn@idiap.ch, pierre.wellner@idiap.ch, philippe.baudrion@adm.unige.ch

Abstract

This paper presents recent results of the application of the task-based Browser Evaluation Test (BET) to meeting browsers, that is, interfaces to multimodal databases of meeting recordings. The tasks were defined by browser-neutral BET observers. Two groups of human subjects used the Transcript-based Query and Browsing interface (TQB), and attempted to solve as many BET tasks – pairs of true/false statements to disambiguate – as possible in a fixed amount of time. Their performance was measured in terms of precision and speed. Results indicate that the browser's annotation-based search functionality is frequently used, in particular the keyword search. A more detailed analysis of each test question for each participant confirms that despite considerable variation across strategies, the use of queries is correlated to successful performance.

1. Introduction

As more multimedia data becomes available, accessing this data and finding relevant information in large multimedia collections requires the design of more powerful search and browsing interfaces. *Meeting browsers* allow users to find potentially relevant information in multimedia archives of meeting recordings, e.g. a series of corporate meetings that were captured in an instrumented meeting room. The goal of this paper is to propose a task-based evaluation method – the BET, for Browser Evaluation Test – and apply to a transcript-based meeting browser, in order to find the most useful features of the browser. The experiment presented here also analyses the overall coherence of the scores in order to assess the validity of the evaluation method itself. Indeed, defining normalized evaluation tasks for meeting browsing is a recent challenge, which, if solved, would allow a uniform comparison of potentially very different multimedia search and browsing technologies.

The evaluation method will be first explained in Section 3. The main features of the Transcript-based Query and Browsing interface (TQB), an annotation-oriented meeting browser, are outlined in Section 4. The particular evaluation setting used in these experiments are described in Section 5. Results and their discussion appear in the last section of the paper, Section 6.

2. Evaluation of interactive software

This section discusses some landmarks in the evaluation of interactive software, especially multi-modal dialogue systems, which is still an open problem (Gibbon et al., 2000; Möller, 2002; Dybkjær et al., 2004). The framework of the ISO/IEC 9126 and 14598 standards for software evaluation suggests that, since the task of meeting browsing does not impose specific requirements, the most appropriate technique is either task-based evaluation, or evaluation in use (ISO/IEC, 2004; Bevan, 2001).

The main parameters to be evaluated are thus *effectiveness* – the extent to which the software helps the user to accomplish a task, *efficiency* – the speed with which the task is accomplished, and *user satisfaction* – measured using questionnaires. A well-known approach to dialogue system evaluation, PARADISE (Walker et al., 1997), predicts user satisfaction from task completion success and from a number of computable parameters related to dialogue cost. The components of a dialogue system can also be evaluated separately using external quality metrics (Traum et al., 2004).

An evaluation task for interactive question answering was proposed in iCLEF, the Interactive track for the Cross-Language Evaluation Forum (Gonzalo et al., 2006), with some important differences with the present work: in our case, the domain is fixed (one meeting), hence the set of possible questions is narrower, and is not defined by the experimenters, but by independent observers; the questions are expressed as true/false alternatives, allowing for automatic scoring, and subjects are scored using precision and speed, and not accuracy alone.

3. The Browser Evaluation Test (BET)

The BET is a framework containing guidelines and software tools that allow evaluators to construct empirically a browser-independent evaluation task (Wellner et al., 2005; Cremers et al., 2006), and then to test the performances of a given browser on that task, as summarized in Figure 1. The BET can be applied independently of the intended specifications of a browser, as it avoids the introduction of developer bias regarding the particular features (such as modalities) that a given browser might implement.

A task consists of a set of *observations of interest* determined by the pool of observers that have watched closely a given meeting recording, and have noted the most salient facts and events that occurred in the meeting. The observations are then sampled and sometimes corrected by the ex-

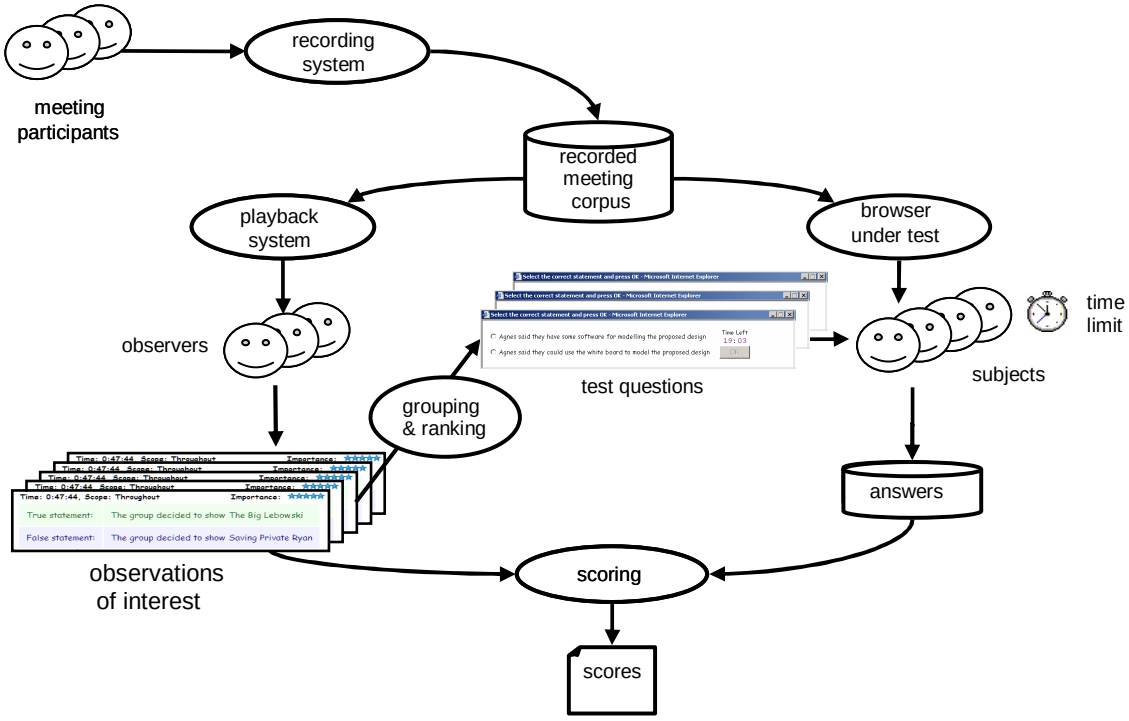

Figure 1: Stages in the design and execution of a BET evaluation.

perimenters to produce a final list of observations. The actual testing of a browser requires subjects to use the browser to review the meeting, and to answer as many test questions as possible in a fixed amount of time (test questions are pairs of true and false statements derived from observations).

Here, BET observations were prepared on two meetings, IB4010 and IS1008c, from the AMI Meeting Corpus (Carletta et al., 2006), in English, involving four speakers each. These meetings were selected quite early in the corpus construction process, as they were among the first to be completely annotated, and their topics were different enough to avoid training effects from the first to the second meeting. Indeed, in the first meeting, the managers of a movie club select the next movie to show, while in the second one, a team discusses the design of a remote control. A third meeting – ISSCO Meeting 024 – which is not part of the official AMI Meeting Corpus as it was not recorded with the same specifications, was used for setup and other experiments; BET observations are also available for this meeting. There are respectively about 130, 60 and 160 pairs of true/false observations, coming from about half-dozen observers, for each meeting.

4. TQB interface for BET experiments

The Transcript-Based Query and Browsing interface (TQB) (Popescu-Belis and Georgescul, 2006) provides access to the transcript of meetings and to their annotations, as well as to the meeting documents, as these language-related modalities are considered to be the main information vector related to human interaction in meetings. TQB is accessible via HTTP, and was thus easily integrated into a larger online application running the BET. The overall layout of the interface is shown in Figure 2, with, in the upper left corner, the BET window containing a true-false pair of statements. Along with the transcript, the following annotations are stored in the database to which the TQB interface gives access: speech segmentation into utterances, dialogue act labeling, thematic episodes labeled with keywords, and document-speech alignment. To avoid errors from automatic recognizers, we use manual transcripts and annotations.

Users of TQB can search for the particular utterances of a given meeting that satisfy a set of constraints on the annotations, including string matching. The results of a query, i.e. the utterances that match all the constraints, are displayed in a separate frame. These utterances can be used as a starting point to browse the meeting, by clicking on one of them, which makes the transcript frame scroll automatically to its position. The transcript frame also provides access to the audio for each utterance.

5. BET setup for TQB experiments

The evaluation proceeds as follows. The subjects were students at the Translation School of the University of Geneva, with no previous involvement in meeting recording or browsing. The subjects first read the instructions for the experiment on a computer screen, which explained first the BET guidelines and then the basic principles of the TQB interface, using a snapshot and 4-5 paragraphs of text. The subjects did not have the opportunity to use TQB before the session, hence the first meeting they saw was their very first occasion to explore its functions.

The BET master interface displays one by one the pairs of true/false statements corresponding to observations. Each subject must determine which one is the true statement (and

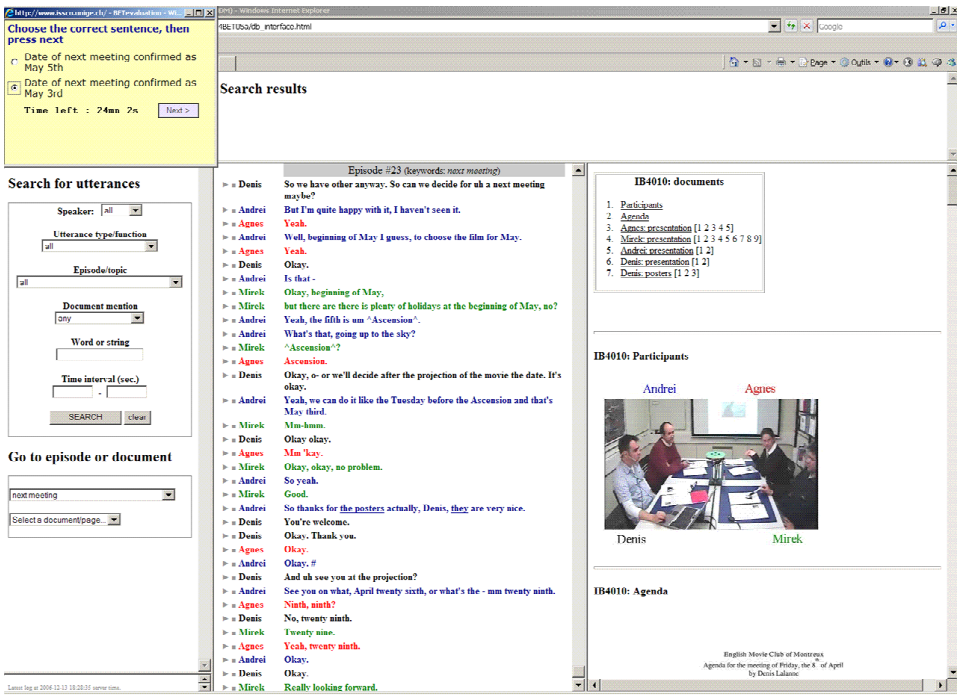

Figure 2: Snapshot of TQB interface with BET window in the upper left corner.

implicitly which one is the false one) by using TQB to browse the meeting. Once the true statement is selected, and the choice validated, the BET interface automatically displays the following pair, and so on until the time allowed for the meeting is over. A pause is allowed, and then each subject proceeds to the second meeting.

Half of the subjects started with IB4010, and then continued with IS1008c, while the other half started with IS1008c followed by IB4010. The duration allowed for each meeting was set at half the duration of the meeting: 24'40" for IB4010 and 12'53" for IS1008c. Overall, results from 28 subjects that have completed both meetings are available.

Two important parameters characterize the subjects' performance. *Precision* is the proportion of correctly solved true/false statements (questions) among all statements that were seen—a number between 0 and 1. *Speed* is the average number of pairs of statements that were processed per unit of time—counted as questions per minute. These scores parallel somewhat precision and recall scores, and are respectively related to effectiveness and efficiency.

6. BET results for the TQB interface

6.1. Overall and meeting-specific scores

The overall precision, averaged for 28 subjects on two meetings, is 0.84 with a ± 0.05 confidence interval at 95% level (95% confidence intervals will be regularly used below). The overall average speed is 0.63 ± 0.09 questions per minute. These values do not vary significantly across the two groups. The average speed and precision vary more markedly across the two meetings, though however these differences are not significant at the 95% confidence level: speed and precision are 0.67 ± 0.10 and respectively 0.85 ± 0.05 for IB4010, both higher than the respective values for IS1008c, 0.57 ± 0.13 and 0.79 ± 0.10. If the statistical significance was higher, one could conclude that IB4010 is easier than IS1008c from the BET perspective.

6.2. Group-specific scores

As the two meetings may have different difficulties, it is safer to compare scores on the same meeting. It is thus possible to compare scores on IB4010 when this meeting is seen first with scores obtained when it is seen second, and similarly for IS1008c.

Figure 3 shows scores of the two groups on meeting IS1008c. The average values of precision and speed are both higher when the meeting is seen in second position, i.e. when the subjects were able to get some training on a previous meeting (here, IB4010). The 95% confidence intervals are however strictly disjoint for precision only. These results point to an important property of the TQB interface: its *learnability*, i.e. the fact that performances on a meeting are generally higher when the subjects have already used TQB on a previous meeting, than when they use TQB for the very first time on that meeting.

6.3. Use of TQB features by the subjects

The analysis of TQB features used during the experiments shows that queries to the transcript and annotation database are quite extensively used to browse meetings. Subjects submit on average 2.50 ± 0.54 queries for each BET question. When subjects use TQB queries, they click on average in 35% of the cases on one or more utterances returned by the query, to visualize them in context. Viewing the utterances within the meeting transcript appears to be sufficient to answer BET questions, as listening to the related audio is very infrequent, only about twice per meeting.

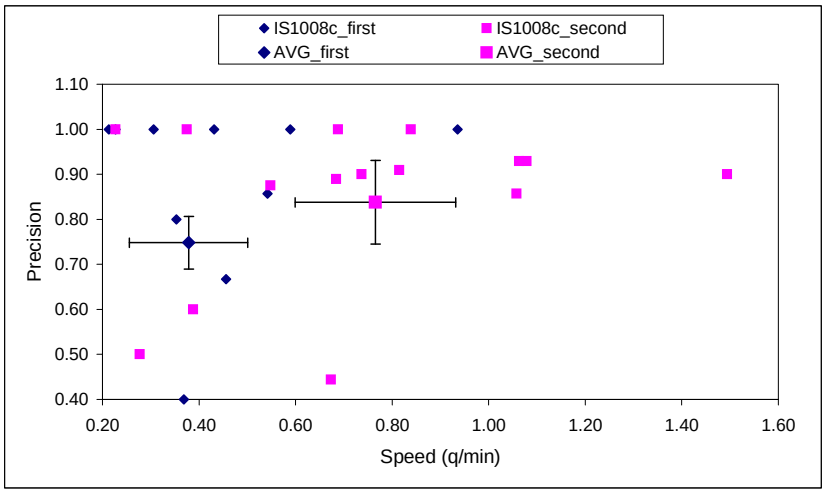

Figure 3: Speed and precision for each subject and average score with 95% confidence intervals on IS1008c. Blue diamonds (◇): subjects seeing the IS1008c meeting first; pink squares (□): subjects seeing it second.

Statistics over all the 550 queries produced by the 28 subjects indicate that most of the queries produced by the subjects, when trying to answer BET questions, are keyword related: 43% of the queries look only for a specified word (or character string), while an additional 31% look for a specific word uttered by a particular speaker, and 7% for a word within a given topic episode. Some other constraints or combinations of constraints are used in 1–3% of the queries each: word(s) + dialogue-act, word(s) + person + dialogue-act, topic, person, word(s) + topic + person, etc. The fact that subjects use the query functionality mainly to do keyword search over the transcripts probably reflects the influence of popular Web search engines, and suggests that annotations other than transcript could better be used for automated meeting processing (e.g. for summarization) rather than directly for search by human users.

6.4. Question-specific scores

It is also possible to compute the above statistics separately for the correct answers, and for the wrong ones, and to compare the results. For instance, the average number of queries per BET question, computed only for the questions to which a subject answered correctly, is 2.41 ± 0.58, while the same average over the wrong answers is 2.01 ± 0.53.

Average values for precision and speed show that while scores generally increase after learning, there is considerable variation across questions, e.g. improvement of precision is not the same for all questions, while speed sometimes even degrades in the second round (for the 4th and 6th questions). These results indicate that performances should be analyzed separately for each question, as their nature requires different competencies and browser functionalities.

7. Conclusion

Overall, the results of applying the BET to the TQB annotation-oriented interface appear to capture a number of properties related to browser quality, which match our a priori intuitions and therefore contribute to the validate the BET evaluation method itself.

The BET offers thus a generic, task-based solution to the problem of evaluating very different meeting browsers, setting few constraints on their functionalities. The set of BET observations created for three meetings will constitute a valuable resource for future evaluations, along with the scores obtained in the experiments presented here, which will provide an initial baseline to which future interfaces can be compared.

Acknowledgments

The authors acknowledge the support of the Swiss National Science Foundation within the IM2 National Center of Competence in Research, and of the European IST Program, within the AMIDA Integrated Project FP6-0033812.

8. References

Nigel Bevan. 2001. International standards for HCI and usability. *International Journal of Human-Computer Studies*, 55:533–552.

Jean Carletta, Simone Ashby, Sbastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction II*, LNCS 3869, pages 28–39. Springer-Verlag, Berlin/Heidelberg.

Anita Cremers, Wilfried Post, Erwin Elling, Betsy van Dijk, Bram van der Wal, Jean Carletta, Mike Flynn, Pierre Wellner, and Simon Tucker. 2006. Meeting browser evaluation report. Deliverable D6.4, AMI Project, December 2006.

Laila Dybkjær, Niels Ole Bernsen, and Wolfgang Minker. 2004. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54.

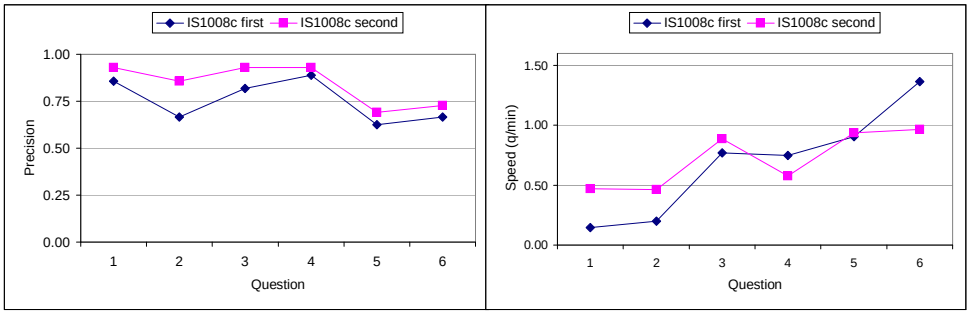

Figure 4: Variations in average speed and precision for each of the first six questions about IS1008c. Blue diamonds (◇): subjects seeing the IS1008c meeting first; pink squares (□): subjects seeing it second.

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors. 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, Dordrecht.

Julio Gonzalo, Paul Clough, and Alessandro Vallin. 2006. Overview of the clef 2005 interactive track. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Mller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories (CLEF 2005 Revised Selected Papers)*, LNCS 4022, pages 251–262. Springer-Verlag, Berlin / Heidelberg.

ISO/IEC. 2004. *ISO/IEC TR 9126-4:2004 (E) – Software Engineering – Product Quality – Part 3: Quality in Use Metrics*. International Organization for Standardization / International Electrotechnical Commission, Geneva.

Sebastian Möller. 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. In *SIGdial 2002 (3rd SIGdial Workshop on Discourse and Dialogue)*, pages 142–153, Philadelphia, PA.

Andrei Popescu-Belis and Maria Georgescul. 2006. TQB: Accessing multimodal data using a transcript-based query and browsing interface. In *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, pages 1560–1565, Genova.

David R. Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, pages 1699–1702, Lisbon.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *ACL/EACL 1997 (35th Annual Meeting of the Association for Computational Linguistics)*, pages 271–280, Madrid.

Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. 2005. A meeting browser evaluation test. In *CHI 2005 (Conference on Human Factors in Computing Systems)*, pages 2021–2024, Portland, OR.