IDIAP RESEARCH REPORT

COMPARING MEETING BROWSERS USING A TASK-BASED EVALUATION METHOD

Andrei Popescu-Belis

Idiap-RR-11-2009

JUNE 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Comparing meeting browsers using a task-based evaluation method

Andrei Popescu-Belis,
Idiap Research Institute, Martigny, Switzerland
andrei.popescu-belis@idiap.ch

June 24, 2009

Abstract

Information access within meeting recordings, potentially transcribed and augmented with other media, is facilitated by the use of meeting browsers. To evaluate their performance through a shared benchmark task, users are asked to discriminate between true and false parallel statements about facts in meetings, using different browsers. This paper offers a review of the results obtained so far with five types of meeting browsers, using similar sets of statements over the same meeting recordings. The results indicate that state-of-the-art speed for true/false question answering is 1.5–2 minutes per question, and precision is 70%–80% (vs. 50% random guess). The use of ASR compared to manual transcripts, or the use of audio signals only, lead to a perceptible though not dramatic decrease in performance scores.

Keywords: meeting browsers, task-based evaluation, spoken information retrieval.

1 Introduction

The increasing amount of multimedia recordings, in particular of human meetings, raises the challenge of accessing the information contained within them. A large variety of component technologies such as speech recognition, diarization, summarization, but also document and video processing, are used to analyze language and other modalities from meeting recordings. Their use has often been justified as facilitating the access to information from recordings, by transforming raw data into more and more abstract layers of representation. In reality, the output of speech/language/multimodal processing technology is often not directly usable by humans for an information access task. Instead, this output must either be rendered via meeting

browsers, or can be used as input for more abstract processing modules[1]. The applicative objectives of meeting processing techniques naturally raise the question of their actual usefulness for a general-purpose information access task.

In this paper, a question answering approach is adopted for the evaluation of tools that enhance access to meeting recordings. The main goal is to discuss several evaluations of meeting browsers that were carried out with similar resources and metrics. These figures provide a snapshot of the state-of-the-art performance for the meeting browsing task at the time of writing, while also illustrating the challenges and variability of task-based evaluation using human subjects.

The paper is organized as follows. Section 2 of the paper introduces the data and the meeting browsers under study. Section 3 describes the Browser Evaluation Test (BET) protocol and its related resources. Section 4 compares the results obtained with several browsers – audio-based, transcript-based, or document-based – evaluated using the BET, in terms of speed and precision, followed by a discussion in Section 5.

2 Meeting Browsers: an Overview

The problems related to meeting recording, processing and retrieval have spun a large body of research in the past decade, and have demonstrated applicative potential as well. The availability of large amounts of transcribed and annotated meeting recordings, e.g. from the ICSI-MR, AMIDA and CHIL projects [1, 2, 3], has allowed numerous studies based on statistical learning, which use the data for training and test. This has also encouraged the development of interfaces and tools called *meeting browsers*, which enable researchers and potential end-users to access the information enclosed in the recordings.

Many scenarios in which meeting browsers answer specific user needs have been described [4, 5, 6, 7], although more user-centric studies of meeting technology are still needed. An important distinction has been made between two types of functions, which both have value, depending on the intended use of the browser: "gisting" is the synthesis of important information contained in meeting recordings (from a certain point of view), while "information access" looks for precise facts located in specific sections of a meeting (which can be part of a large collection). Both types of functions can be accomplished either over one meeting, or across several meetings, and can use features extracted from any combination of modalities available in the recordings. In most cases, however, the spoken language modality plays a central role, either as audio or ASR transcript; manual transcripts

[1]For instance, automatic summarization of meetings can use the output of speech recognizers along with utterance segmentation and dialogue act recognition.

are mainly used for testing purposes.

In this paper, we focus on the information access task over a given meeting (as opposed to tasks that require an abstraction over an entire meeting). The goal is to locate specific bits of information within a meeting that typically lasts between 30 minutes and one hour. A large number of research-oriented browsers were designed for this type of task [7, 8], using various types of features extracted from multimedia recordings: speech, transcript, annotations, documents, and videos.

Most of the browsers quoted in this paper are outlined in [8], apart from the audio-based browsers described in [9]. Each browser renders a different subset of media/modalities of the meeting recordings, and offers different search criteria. The *audio-based browsers* use automatic processing to provide access to audio, speaker segmentation and slides, with two possible techniques: (a) speedup audio, based on processing that avoids the chipmunk effect; and (b) overlap audio, playing two different parts of a meeting in the left vs. right channel, assuming that the user will take advantage of the cocktail party effect to focus on only one of them at a time.

The *JFerret* browser [10, 11], in fact a sample implementation of the JFerret framework, offers access to audio, video, slides, ASR transcript, and a temporal representation of speaker segmentation (see also [7, Section 5]). The *Transcript-based Query Browser, TQB* [12, 13], uses a number of manual annotations and compares their respective merits when using the browser: apart from manual transcript, audio and slides, it also contains dialogue act annotation and topic labeling. *JFriDoc* [14] is a document-centric browser that exploits the alignments between documents and speech.

Finally, *Archivus* [15, 16] is a multimodal browser that uses a reference transcript with annotations, and allows users to engage in spoken or written dialogue with the browser in order to set the search parameters and obtain results. However, human operators are required behind the scenes to run the dialogue and ASR engine, therefore the Archivus evaluation is in reality a Wizard-of-Oz experiment aimed at obtaining feedback on modality use.

3 Evaluation of Meeting Browsers

3.1 Overview of Existing Methods

The evaluation of interactive software, especially of multi-modal dialogue systems, is still an open problem [17, 18]. As the task of meeting browsing does not impose specific functionality requirements that can be tested separately, the most appropriate technique appears to be task-based evaluation in use. The main quality aspects to be evaluated are thus *effectiveness* – i.e. the extent to which the software helps the user to accomplish a task, *efficiency* – i.e. the speed with which the task is accomplished, and *user satisfaction*, which is measured using questionnaires. In fact, a well-known

approach to dialogue system evaluation, PARADISE [19], has shown that user satisfaction stems from task completion success and from dialogue cost, therefore one should focus indeed on effectiveness and efficiency. Approaches advocating task-based evaluation of meeting browsers on very specific tasks were proposed in [7] and [20].

Systematic evaluation of QA systems has started with the TREC-8 QA task in 1999 [21, 22]. The campaign devised an original procedure to obtain non-biased questions, using multiple sources (participants, assessors, organizers, and one Web-based QA system), and leading to a set of 1,337 questions, of which 200 were selected for the campaign. At TREC 2003, the test set of questions contained 413 questions (3 types: factoid, list, definition) drawn from AOL and MSNSearch logs [23]. An evaluation task for interactive QA was also proposed at iCLEF, the Interactive track of the Cross-Language Evaluation Forum [24]; systems-plus-humans where evaluated for accuracy over a large set of questions defined by the experimenters.

3.2 The Browser Evaluation Test (BET)

The Browser Evaluation Test, or BET [11, 13], is a procedure to collect browser-independent "questions" and to use them for evaluating a browser's capacity to help human users answering them. These questions are in fact pairs of parallel true/false statements (see examples below) which are constructed by neutral "observers" that view a meeting and first write down "observations of interests" about the meeting, then create the false counterpart of each statement. As several observers make observations for a given meeting, the results are put together by experimenters which gather similar observations into groups, and choose only one representative for each group. The importance of the group is derived from the observers' rating of importance and the size of each group.

Three meetings from the AMI Corpus [2], in English, were selected for the observation collection procedure: IB4010, IS1008c, and ISSCO-Meeting_024. For these meetings, respectively 222, 133 and 217 raw observations were collected, from 9, 6 and 6 observers, resulting in 129, 58 and 158 final pairs of true/false observations. These figures are in the same range as those from the TREC QA campaigns, though the data set is here considerably smaller (one meeting vs. a large collection of documents). The average size of the similarity groups was found to be 1.72, 2.29 and 1.37 observations per group. As a measure of inter-observer agreement, these values are not very high, but they are much higher for the observations ranked highest for importance, and which are therefore shown in priority to subjects. Therefore, the agreement for the first 16 observations on IB4010 and the first 8 on IS1008c is respectively 55% and 83%. As an example, the two most important observations (pairs of statements) for the IB4010 meeting are in Table 1.

Table 1: First three most quoted observations of interests for the two meetings, represented as pairs of true and false statements.

	Movie club meeting (IB4010)	**Remote control design meeting (IS1008c)**
TRUE	The group decided to show The Big Lebowski.	According to the manufacturers, the casing has to be made out of wood.
FALSE	The group decided to show Saving Private Ryan.	According to the manufacturers, the casing has to be made out of rubber.
TRUE	Date of next meeting confirmed as May 3rd.	Christine is considering cheaper manufacture in "other countries" before backtracking and suggesting the remote could support a premium price [...].
FALSE	Date of next meeting confirmed as May 5th.	Ed is considering cheaper manufacture in "other countries" before backtracking and suggesting the remote could support a premium price [...].
TRUE	Denis informed the team that the first objective was to choose a film and the second was to discuss an advertising poster.	The product is expected to last over several hundred years.
FALSE	Denis informed the team that the first objective was to choose a film and the second was to discuss a date for the film to be shown.	The product is expected to last more than 5 but less than 15 years.

Table 2: *Comparative results of several meeting browsers evaluated in similar conditions (see text for actual differences), in terms of average time needed by subjects to answer a question, and of average precision. Standard deviations (or confidence intervals at 95% when marked with a "*") are in absolute (not relative) values of time or precision.*

Browser	**Condition**	**Nb. of subjects**	**Time per question (s)**	**Stdev***	**Precision**	**Stdev***
Audio-based	Speedup	12	99	26*	0.78	0.06*
browsers	Overlap	15	88	23*	0.73	0.08*
JFerret	BET set (pilot)	10	100	43	0.68	0.22
sample	Gisting (5 questions)	5	max. 180	0	0.45	0.34
	Factual (5 questions)	5	max. 180	0	0.76	0.25
Transcript-	1st meeting	28	228	129*	0.80	0.09*
based browser	2nd meeting	28	92	16*	0.85	0.06*
(TQB)	Average on both meetings	28	160	66*	0.82	0.06*
Document-	With speech/document links	8	113	n/a	0.76	n/a
based (FriDoc)	Without links	8	136	n/a	0.66	n/a
Archivus	T/F questions	80	127	36	0.87	0.12
multimodal	Open questions	80	==	==	0.65	0.22

To apply the BET questions, subjects are required to view the pairs of BET statements in sequence and decide, using the meeting browser, which statement from the pair is true, and which one is false. These pairs are presented in decreasing order of importance, and the ordering is checked so that previous questions do not give away the answers to subsequent ones.

The time allowed for each meeting is half the duration of the meeting, i.e. 24'40" for IB4010, and 12'53" for IS1008c, and the subjects are shown new questions as long as the allowed time is not over. Another approach that was alternatively adopted is to fix the number of questions and leave the subjects at their own pace, with a large upper bound to avoid subjects watching an entire meeting before answering.

Apart from observing the subjects behavior with the browser, and looking at their satisfaction through post-experiment questionnaires, two scores are generally computed, viz., precision and speed[2]. One should compute the average time to answer a question (per question, per group, or both) rather than the average speed, because time is an additive quantity, but not speed[3].

4 BET Evaluation of Meeting Browsers

The BET resource was used to evaluate a number of browsers, and available results are compared in this section. Ideally, of course, such a comparison is licensed only if the same questions were used, in the same order, on comparable groups of subjects, trained in similar conditions, and having the same amount of time at their disposal[4]. These conditions are rarely met, except in strictly controlled evaluation campaigns, which have never been organized yet for meeting browsers. Until such a campaign is set up, the only possible comparison is the one attempted here, which lists and discusses the main evaluation details for each browser, in the same order as in Table 2.

The *audio-based browsers* – speedup and overlap – were evaluated with a group of native English speakers at the U. of Sheffield [9]. (A "base" browser was also evaluated but we focus here on the audio-based ones.) Subjects were trained first to answer BET questions using a simple player on one meeting (ISSCO-Meeting_024), then answered the "official list" of BET questions using either of the audio-based browsers, on one of the two remaining meeting drawn randomly, in 50% of meeting time. Usable results were obtained from a dozen subjects per condition, though with large

[2]Precision is related to "effectiveness" and speed is related to "efficiency".

[3]The average of speeds is typically not just the arithmetic average of speed values, but should be calculated instead using the average time. For instance, The figures for audio-based browsers are highly biased in [9] because they average speeds – with times, the performances appear to be quite lower.

[4]A group cannot be used more than once because of the training effect, given the small amount of meetings that are available.

variability across subjects. For instance, on the overlap browser, two subjects answered questions about three times faster than average, making one wonder whether they simply gave up trying to find the correct answer and answered almost randomly (their total precision is 0.64 and 0.72). If these subjects are removed from the data, average time increases from 88 to 98 seconds, and average precision from 0.73 to 0.74 (a very small increase due to the fact that precision is overall quite low).

A sample *JFerret* browser was evaluated as a trial run in the original BET paper [11], and re-tested later with different questions in [7, p. 210-211]. In the trial run, 10 subjects answered questions from the "official set" over ISSCO-Meeting_024 in half the meeting's duration. In the second run, five subjects (from the U. of Sheffield) answered five BET-inspired factual questions, as well as five questions that required gisting over the entire meeting; answers were in open form, not binary; and time was limited to 30 minutes, though apparently none of the subjects used the entire interval. None of the conditions allowed any training before the trial.

The *TQB* browser was evaluated with 28 subjects from the U. of Geneva [13], half of which started with the IS1008c meeting and proceeded with the IB4010, the other half doing the reverse. The evaluation used the "official" BET set and allowed 50% of meeting time. The difference in conditions allowed a measure of the training effect over one meeting, but also showed that meetings and related questions are not equally difficult. The experiment found that subjects tend to focus on keyword search over the transcript, sometimes adding constraints on the speaker's name.

The *JFriDoc* document-based browser [14] was also tested using BET-inspired questions, to assess the merits of speech/document links by comparing two conditions: with vs. without enabled links. Eight students from the U. of Fribourg, Switzerland, had to answer 12 questions in each condition (on different meetings), and were allowed at most 3 minutes per question, but exact time was computed too.

Finally, the *Archivus* interactive multimodal browser was tested in a Wizard-of-Oz environment, using a particularly large number of subjects (ca. 80 from U. of Geneva and EPFL) who tested various combinations of modalities. Subjects had to answer 20 questions in 20 minutes, in two conditions (first with a subset of all modalities, then with all of them – we use here results from the second one only). These included true/false questions and short-answer questions, in BET-style. Archivus gave access to a database of six meetings instead of only one [16, chapter 6.6]. The results reported here are separated according to the type of expected answers (note that "short answer" questions do not have a 50% baseline score) though average time is only known jointly. The average time shown in Table 2 includes the system's response time (as for all other browsers), which is on average 36 seconds! This delay is due to the presence of human "wizards" which interpret the user's actions to generate the system's ones. For fairness

to the other browsers, we give the total time, but user time is in reality 36 seconds shorter.

5 Discussion

Comparison across the speed and precision results presented in Table 2 must be taken with a grain a salt, given that baselines are not all at 50% (random binary choice), timing is not always constrained in a similar way, and training conditions and subjects differ across experiments. Moreover, the amount of knowledge provided to the browsers varies greatly, from fully automatic pre-processing of the meetings to manual annotation or human wizards behind the scenes. Therefore, the point here is not to find "the best browser", but rather to provide a range of scores that can be used for future comparison.

Average time per question varies from about 1.5 minutes up to 4 minutes (with no prior training); most browsers-plus-subjects require on average ca. 2 minutes to answer a question. Any significant improvement to this value would be welcome. However, as quick answers are sometimes due to bored subjects who give up searching when questions seem too difficult, a method to detect this strategy would be welcome too. The observed standard deviations for speed are quite high in comparison with those for precision, illustrating the large variability of human speed on this type of task – a challenge when two conditions must be reliably compared.

Precision – generally against a 50% baseline except for some conditions of JFerret and Archivus – generally stays in the 70–80% range, with highest values around 85% (TQB) and 87% (Archivus). This observation suggests that more knowledge is marginally helpful to increase precision, though this often means that subjects spend slightly more time to actually look for the right answer. The STDEVs are somewhat smaller than for speed, but individual performance still varies substantially across subjects.

It is remarkable that all browsers score above the 65% mark (for the binary decision task), which means that the difficulty of the BET is not as high as expected. It might for instance be the case that observers, despite the instructions, did not manage to make the false counterparts of the statements not easily guessable. Detailed analyses for each browser should be able to determine if a subject's answer was indeed supported by the data that they browsed at that time; but if they use previously seen data, this will be much harder to detect from a browser's logs.

6 Conclusion

This paper gathered the available results of meeting browser evaluation using a standardized resource, the BET. The numeric results show that bi-

nary discrimination of true/false statements by humans assisted by browsers generally has 70-80% precision, with peaks reaching 90% if accurate transcripts/annotations are present. However, each question requires on average at least 1.5 minutes to be solved, and often more, up to 2-4 minutes on average for some subjects.

The difficulty of comparing the different results lies in the possible variety of experimental conditions, which could be overcome only in a totally uniform evaluation campaign, which would require a large pool of comparable human subjects, and more resources, i.e. meetings and related questions. In any case, more such resources are a priority for the future – but the existence of the current set already enabled an initial assessment of browsing performance.

7 Acknowledgments

This work has been supported by the Swiss National Science Foundation, through the IM2 National Center of Competence in Research, and by the European IST Program, through the AMIDA Integrated Project FP6-0033812.

References

[1] A. Janin, D. Baron, J. A. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *ICASSP 2003*, Hong Kong, 2003.

[2] J. Carletta *et al.*, "The AMI Meeting Corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction II*, ser. LNCS 3869, S. Renals and S. Bengio, Eds. Berlin: Springer, 2006, pp. 28–39.

[3] D. Mostefa *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3, pp. 389–407, 2007.

[4] S. Tucker and S. Whittaker, "Accessing multimodal meeting data: Systems, problems and possibilities," in *Machine Learning for Multimodal Interaction I*, ser. LNCS 3361, S. Bengio and H. Bourlard, Eds. Berlin: Springer, 2005, pp. 1–11.

[5] A. Lisowska, A. Popescu-Belis, and S. Armstrong, "User query analysis for the specification and evaluation of a dialogue processing and retrieval system," in *LREC 2004*, Lisbon, 2004, pp. 993–996.

[6] A. Nijholt, R. Rienks, J. Zwiers, and D. Reidsma, "Online and off-line visualization of meeting information and meeting support," *The Visual Computer*, vol. 22, no. 12, pp. 965–976, 2006.

[7] S. Whittaker, S. Tucker, K. Swampillai, and R. Laban, "Design and evaluation of systems to support interaction capture and retrieval," *Personal and Ubiquitous Computing*, vol. 12, no. 3, pp. 197–221, 2008.

[8] D. Lalanne *et al.*, "The IM2 multimodal meeting browser family," IM2 NCCR, Technical Report, March 2005. [Online]. Available: http://diuf.unifr.ch/im2/meetingbrowser/IM2BrowserReportMarch2005.pdf

[9] AMI, "Meeting browser evaluation report," AMI EU Integrated Project, Deliverable D6.4, December 2006. [Online]. Available: http://www.amiproject.org/ami-scientific-portal/documentation/annual-reports/pdf/D6_4.pdf

[10] P. Wellner, M. Flynn, and M. Guillemot, "Browsing recorded meetings with Ferret," in *Machine Learning for Multimodal Interaction I*, ser. LNCS 3361, S. Bengio and H. Bourlard, Eds. Berlin: Springer, 2005, pp. 12–21.

[11] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker, "A meeting browser evaluation test," in *CHI 2005*, Portland, OR, 2005, pp. 2021–2024.

[12] A. Popescu-Belis and M. Georgescul, "TQB: Accessing multimodal data using a transcript-based query and browsing interface," in *LREC 2006*, Genova, 2006, pp. 1560–1565.

[13] A. Popescu-Belis, P. Baudrion, M. Flynn, and P. Wellner, "Towards an objective test for meeting browsers: the BET4TQB pilot experiment," in *Machine Learning for Multimodal Interaction IV*, ser. LNCS 4892, A. Popescu-Belis, H. Bourlard, and S. Renals, Eds. Berlin: Springer, 2008, pp. 108–119.

[14] M. Rigamonti, D. Lalanne, F. Evequoz, and R. Ingold, "Browsing multimedia archives through intra- and multimodal cross-documents links," in *Machine Learning for Multimodal Interaction II*, ser. LNCS 3869, S. Renals and S. Bengio, Eds. Berlin: Springer, 2006, pp. 114–125.

[15] M. Ailomaa, M. Melichar, M. Rajman, A. Lisowska, and S. Armstrong, "Archivus: a multimodal system for multimedia meeting browsing and retrieval," in *COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, 2006, pp. 49–52.

[16] M. Melichar, "Design of multimodal dialogue-based systems," PhD Thesis, EPF Lausanne, 2008.

[17] D. Gibbon, I. Mertins, and R. K. Moore, Eds., *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer, 2000.

[18] L. Dybkjær, N. O. Bernsen, and W. Minker, "Evaluation and usability of multimodal spoken language dialogue systems," *Speech Communication*, vol. 43, no. 1–2, pp. 33–54, 2004.

[19] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *ACL/EACL 1997*, Madrid, 1997, pp. 271–280.

[20] W. M. Post, E. Elling, A. H. M. Cremers, and W. Kraaij, "Experimental comparison of multimodal meeting browsers," in *HCII 2007*, Beijing, 2007.

[21] E. M. Voorhees and D. Tice, "The TREC-8 question answering track evaluation," in *TREC-8*, ser. NIST Special Publication 500-246, Gaithersburg, MD, 1999, pp. 83–106.

[22] E. M. Voorhees, "The TREC question answering track," *Natural Language Engineering*, vol. 7, no. 4, pp. 361–378, 2001.

[23] ——, "Overview of the TREC 2003 question answering track," in *TREC 2003*, ser. NIST Special Publication 500-255, Gaithersburg, MD, 2003, pp. 54–68.

[24] J. Gonzalo, P. Clough, and A. Vallin, "Overview of the CLEF 2005 interactive track," in *Accessing Multilingual Information Repositories (CLEF 2005 Revised Selected Papers)*, ser. LNCS 4022, C. Peters *et al.*, Eds. Berlin: Springer, 2006, pp. 251–262.