Human Language Technology: Applications to Information Access

Lesson 11: Meeting Browsers

December 22, 2016

EPFL Doctoral Course EE-724
Andrei Popescu-Belis, Idiap Research Institute

The problem

- What can we do to help people find information in archives of multimedia meeting recordings?

- Alternative answers
  1. First find out what people need, then design and implement
  2. First show people what is possible (design and implement), then find out if they need/like it
  3. Try 1 → 2 → 1 → 2 → …

Meeting browsers: a definition

- Assistance tools that help humans navigate through multimedia records of meetings

- Help people to achieve two goals
  1. Get a general idea about a meeting's content
  2. Find specific pieces of information in meetings
     - either previously unknown to the user (*discovery*)
     - or already known but uncertain (*verification*)

Plan of the lesson

- Outline
  - software design for HLT applications (including meeting browsers)
  - extracting user needs for m. b.
  - designing multimedia m. b.
  - evaluating m. b. in use

- Note
  - this work is related to the achievements and lessons learned from three large projects: Swiss IM2 (2002-2013) and EU AMI + AMIDA (2004-2010)

Software development process

- Waterfall model
  - users formulate requirements (needs) for a task
  - designers write specifications based on them
  - developers create a product that satisfies specifications
  - the product is evaluated against specifications and task
- Difficulties of this model for HLT
  - users' needs are often *underspecified* **or** *beyond reach*
  - designers may also suggest useful functionalities
- Solution: iterative development
  - back-and-forth exchanges between users and developers

Meeting support technology: two methods to elicit user requirements

1. Look at how people use existing technology in order to infer new needs (requirements)
   - good for assessing current practice
   - but how to infer precise specifications for technology that does not exist yet?
2. Ask users to describe functionalities that would "help them with meetings"
   - users must be guided towards a task based on what is feasible → possible bias
     - if not guided, suggestions may be totally unrealistic

User studies for meeting support technology

Jaimes [7]	15	interviews	practice	Importance of audio-visual records for checking or better understanding specific points in a meeting.
	519	questionnaires	practice	Importance of visual cues for recall.
Whittaker [21]	12	interviews	practice	Importance of personal notes, need for to-do summaries.
Cremers [4]	8	interviews	practice / needs	Need for summaries and to-do lists.
Bertini [2]	118	questionnaires	practice / needs	Low utility of audio-visual records, except for persons who missed a meeting or for finding specific information.
Banerjee [1]	12	interviews	practice	Importance of thematic content.
Lisowska [10]	28	elicitation of queries	needs	Heterogeneity of queries, either about the interaction or about simple items in meetings.
Wellner [20]	21	elicitation of observations of interest	needs	Importance of facts, decisions and arguments leading to them, agenda, and dates.
Lisowska [9]	91	Wizard-of-Oz	needs	Importance of training for modality choice in meeting browsing.

Synthesis of user studies (1)

- User requirements vary a lot across studies

- Main dimensions of user requirements
  1. Targeted time span: utterance, fragment, meeting
  2. Targeted media: audio, video, docs, slides, emails
  3. Complexity of searched information: present in the media or inferred from content
  4. Complexity and modality of query

- Depending on context, the expressed needs cover each possible value of each dimension (!)

Synthesis of user studies (2)

- Entire recordings are seen as useless without tools enabling “intelligent” access to their content
- Two types of tools
  1. Summary of an entire meeting
  2. Detailed information related to a meeting
     a. “easy” to extract from metadata and files
        – dates, participants, documents, presentations
     b. “difficult”, requires some form of content analysis
        – decisions and tasks; other facts and arguments; aspects of interaction or media; agenda; date of next meeting

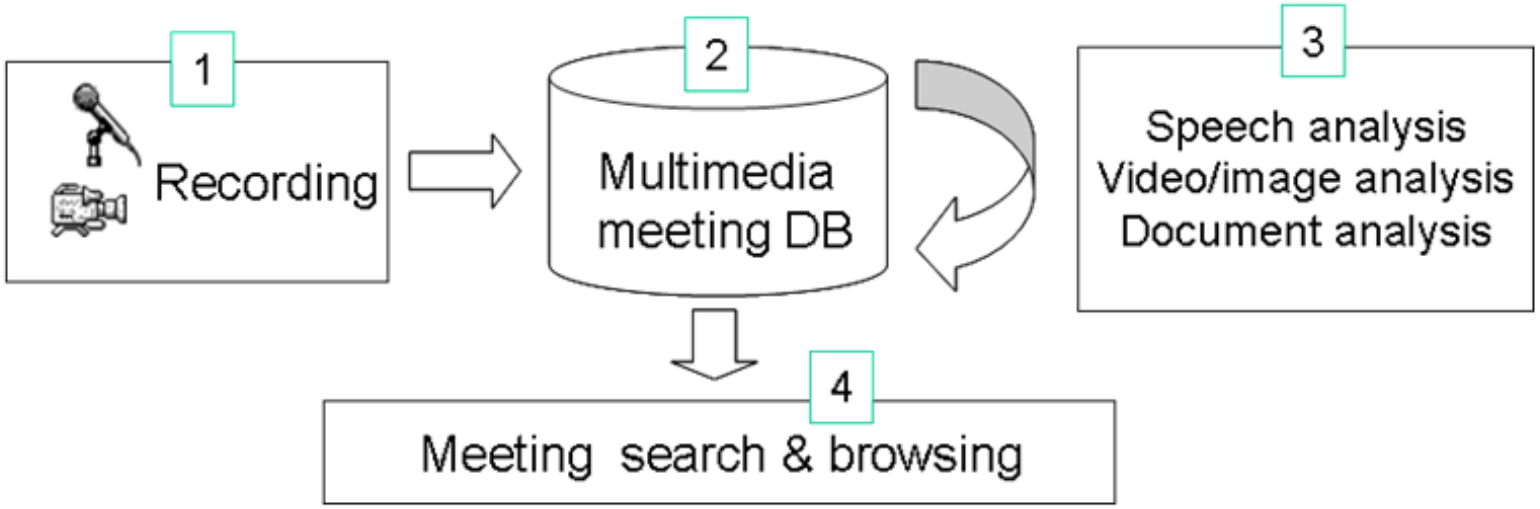
→ Two main applications: summarizers & browsers

Examples of both types

1. Meeting summarization systems
   - structured around its main topics (CMU ISL “Meeting Browser”)
   - structured around the action items / tasks (CALO browser)
2. Fact finding or verification
   - check figures, decisions, assigned tasks, document fragments
   - analyze meeting data to build high-level indexes
     - features: speech transcript, turn taking, attention focus, slides, notes
   - integrated in multimodal interfaces → locate information

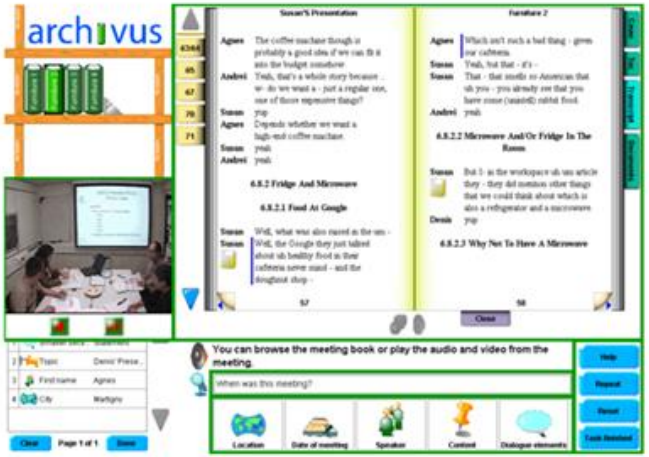
- Surveys
  - M.M. Bouamrane and S. Luz, “Meeting Browsing: State-of-the-Art Review”, *Multimedia Systems*, 12:45, 2007.
  - S. Tucker and S. Whittaker, “Accessing Multimodal Meeting Data: Systems, Problems, and Possibilities”, *Machine Learning for Multimodal Interaction*, LNCS 3361, Springer-Verlag, 2005.
  - Z. Yu and Y. Nakamura, ‘‘Smart Meeting Systems: A Survey of State-of-the-Art and Open Issues,‘‘ ACM Computing Surveys, 42:2, 2010.

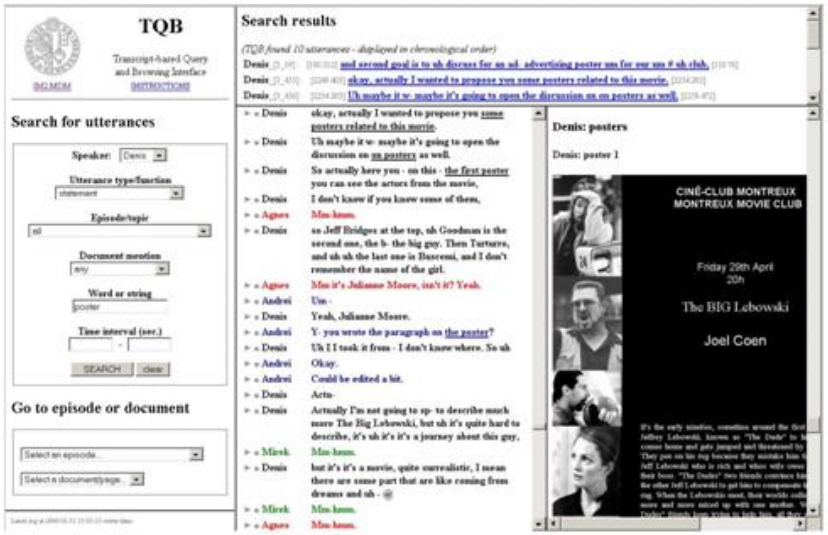
Meeting browsers for fact finding



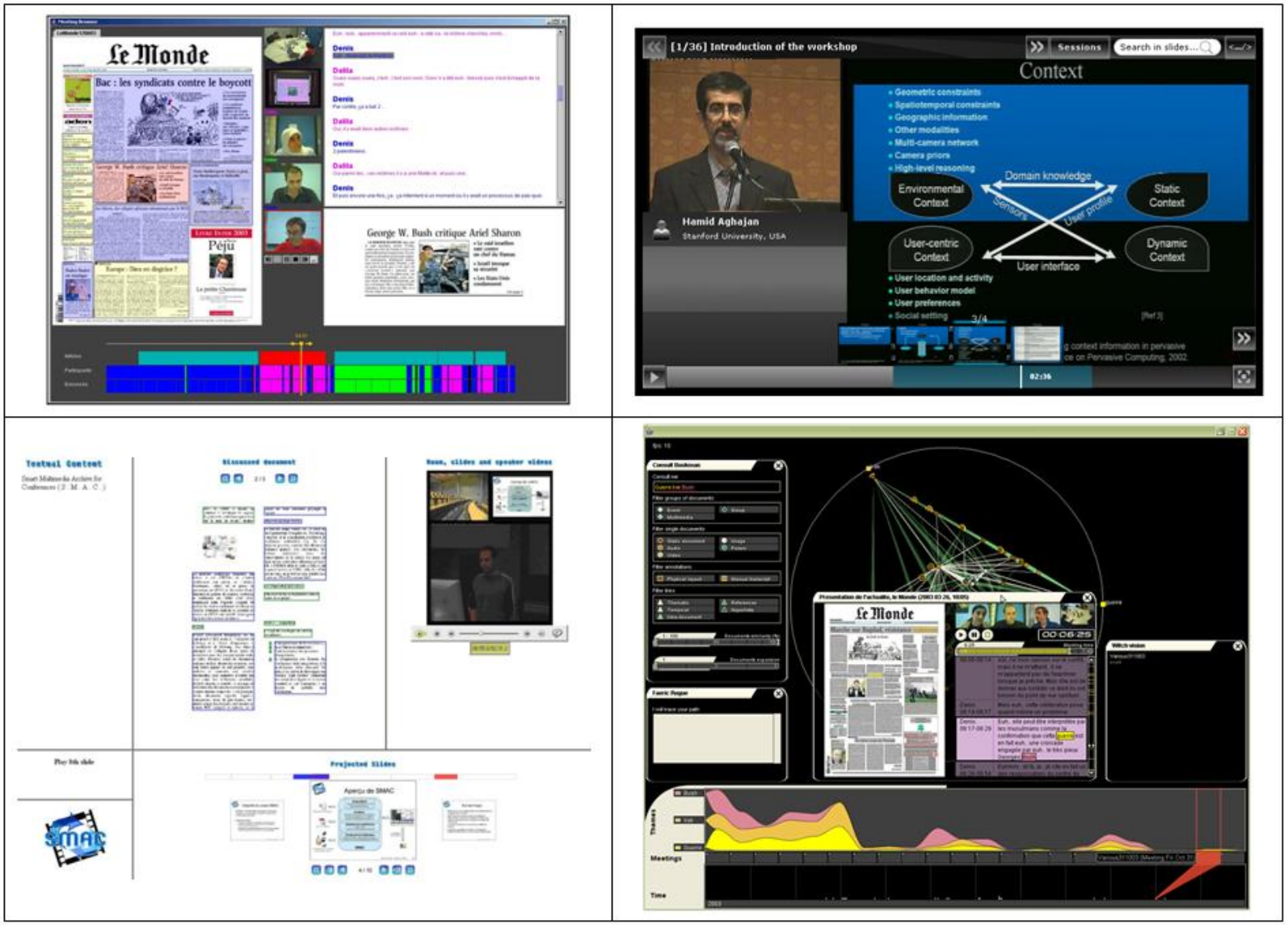
- Speech-centric browsers
  - use audio recordings and/or the transcript
  - often with video
  - sometimes with higher-level annotations
    - named entities, thematic episodes, keywords, etc.
- Document-centric browsers
  - use content of documents related to meetings
  - sometimes with annotations
    - slide change, speech/ document alignment

Examples of speech-centric browsers





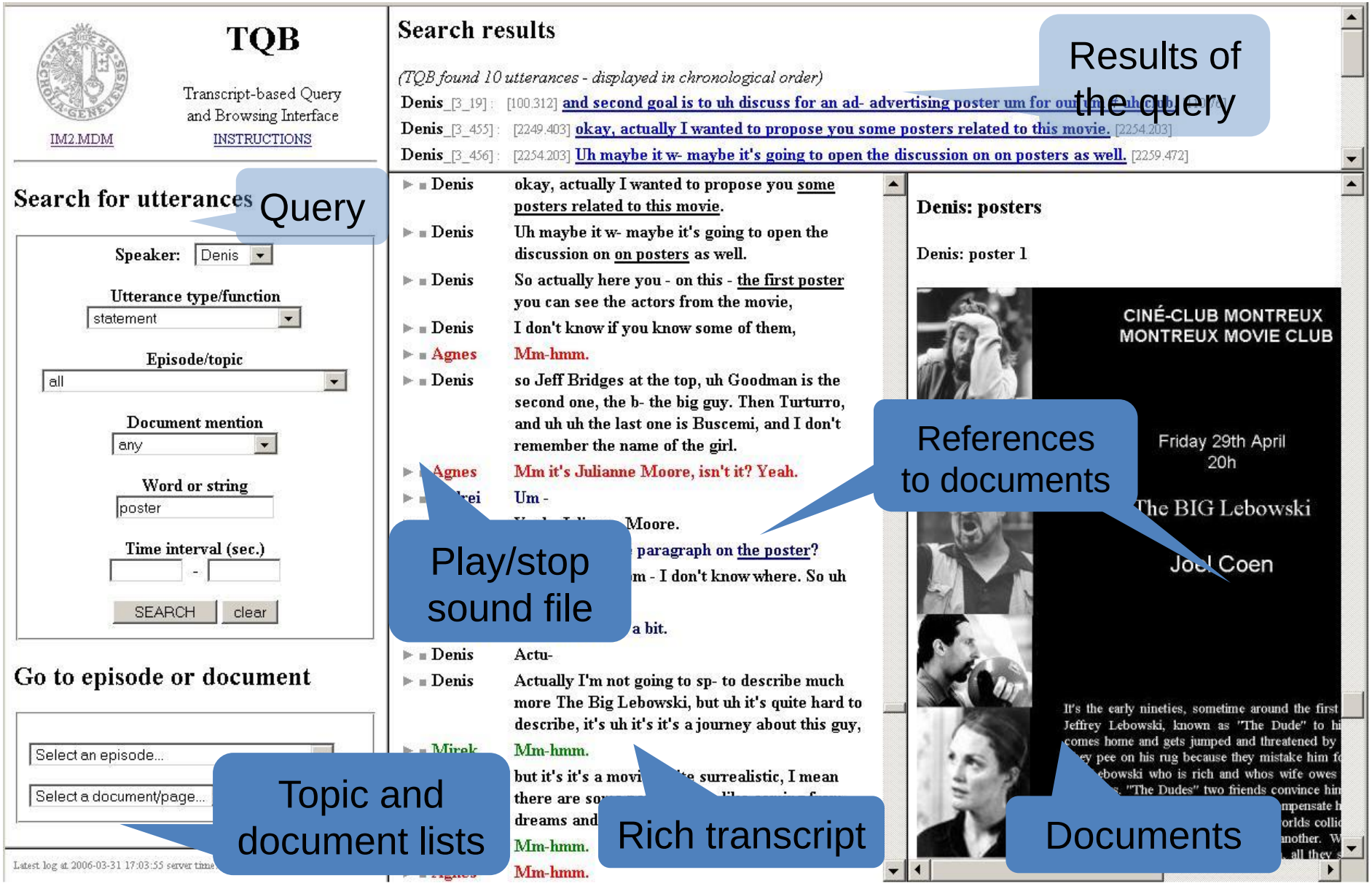
Examples of document-centric browsers



A sample meeting browser: TQB the Transcript-based Query & Browsing interface

- Available media and annotations
  - audio, documents (slides, notes), snapshot of room, but no video
  - manual transcript aligned with audio track
  - utterance segmentation, dialogue acts
  - topic segmentation, keywords, references to documents
- Note: TQB can also use ground-truth annotations and transcript in order to test the impact of imperfect processing
- Using TQB
  - users can query each of the above annotations
    - possibly values for each field are displayed
  - TQB returns all utterances
  - each result can be viewed in its meeting context (transcript + audio)

TQB example : looking for *statements about "poster" by "Denis"*



Evaluation of meeting browsers: the BET protocol

How to evaluate a meeting browser?

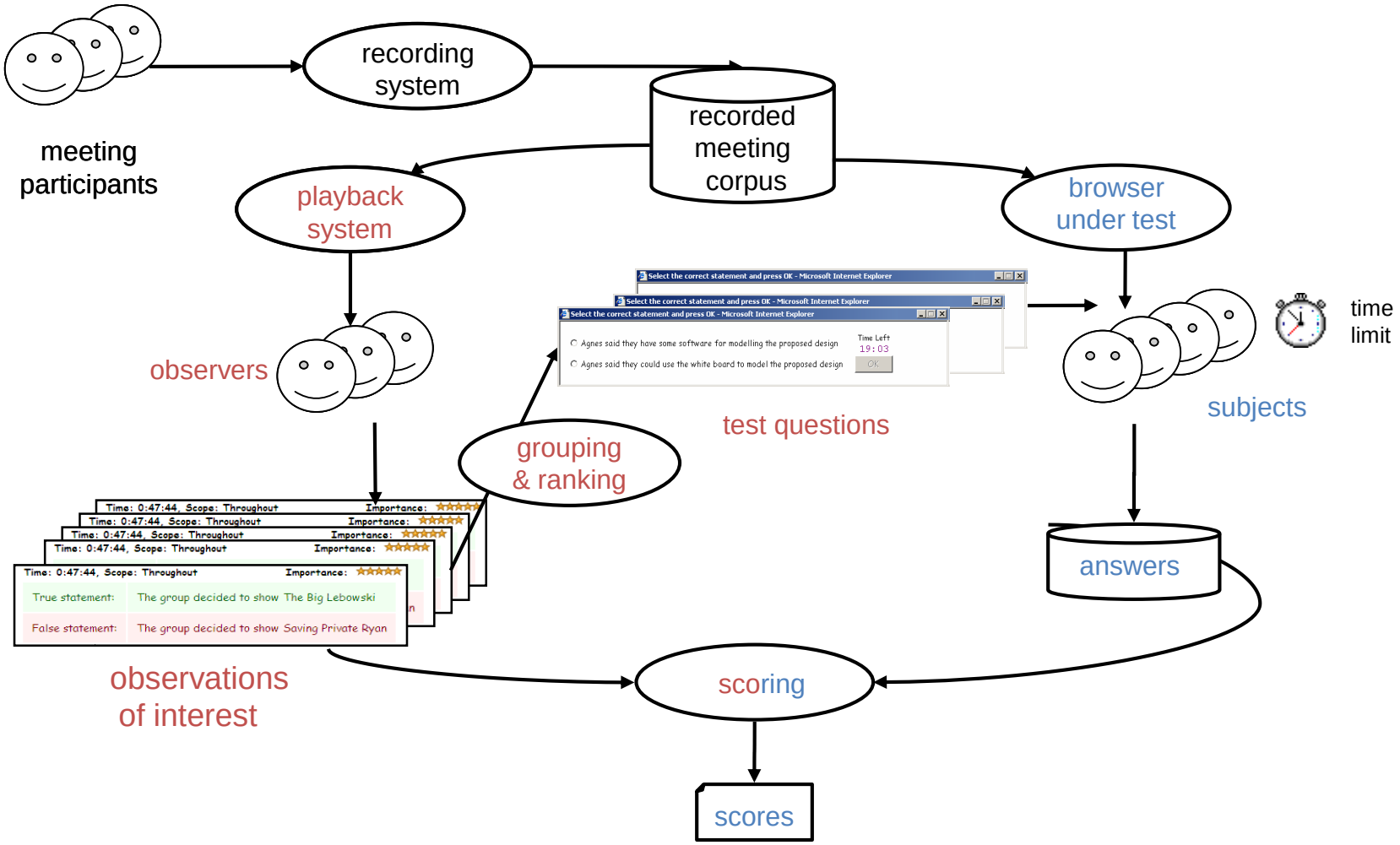
- TREC Question Answering task (≥ TREC-8, 1999)
  - provides series of test questions and correct answers
  - evaluation of fully automated QA systems:
    - similarity of strings AND correctness of supporting document
- Who defined the questions?
  - TREC QA combined submissions from all participants
- Adaptation to meeting browser evaluation
  - ask “neutral” observers to define questions
  - evaluate humans who are using meeting browsers

The Browser Evaluation Test

1. Collect “questions” about a meeting
   - observers view a meeting recording
   - formulate pairs of parallel statements about it
     - observations of interest = facts that were salient for participants
     - one statement is factually true, the other is false
   - rank statements based on importance (# of observers)
2. Use a browser to answer “questions” in limited time
   - i.e. subjects must discriminate T vs. F in BET pairs
3. Measure performance
   - precision (# of correctly discriminated pairs) ~ effectiveness
   - speed (# of pairs processed per unit of time) ~ efficiency

Outline of BET definition & application

(from Wellner et al. 2005)



The BET test set

- 3 meetings from AMI
  - IB4010: movie club
  - IS1008c: remote control
  - ISSCO-024: furnishing
- 21 observers
- 572 pairs of statements
  - consolidated into 350 pairs
  - average size of consolidated groups
    - ~2 for all groups
    - ~5 for the questions used
    - this is a measure of "inter-observer" agreement on what facts are important

- Scope of statements
  - 63% refer to specific moments in a meeting
  - 30% refer to short intervals
  - 7% about entire meeting
- Content of statements
  - decisions (8%)
  - other stated facts, including arguments (76%)
  - related to the interaction or the media (11%)
  - about the agenda (2%)
  - date of next meeting (2%)

Sample questions: T/F pairs

- IB4010 – Movie Club
  - The group decided to show The Big Lebowski /// The group decided to show Saving Private Ryan
  - Agnes did not like the third advertising poster, it had too many colours /// Agnes did not like the third advertising poster, it had no colour
  - Everyone had seen Goodfellas /// No one had seen Goodfellas

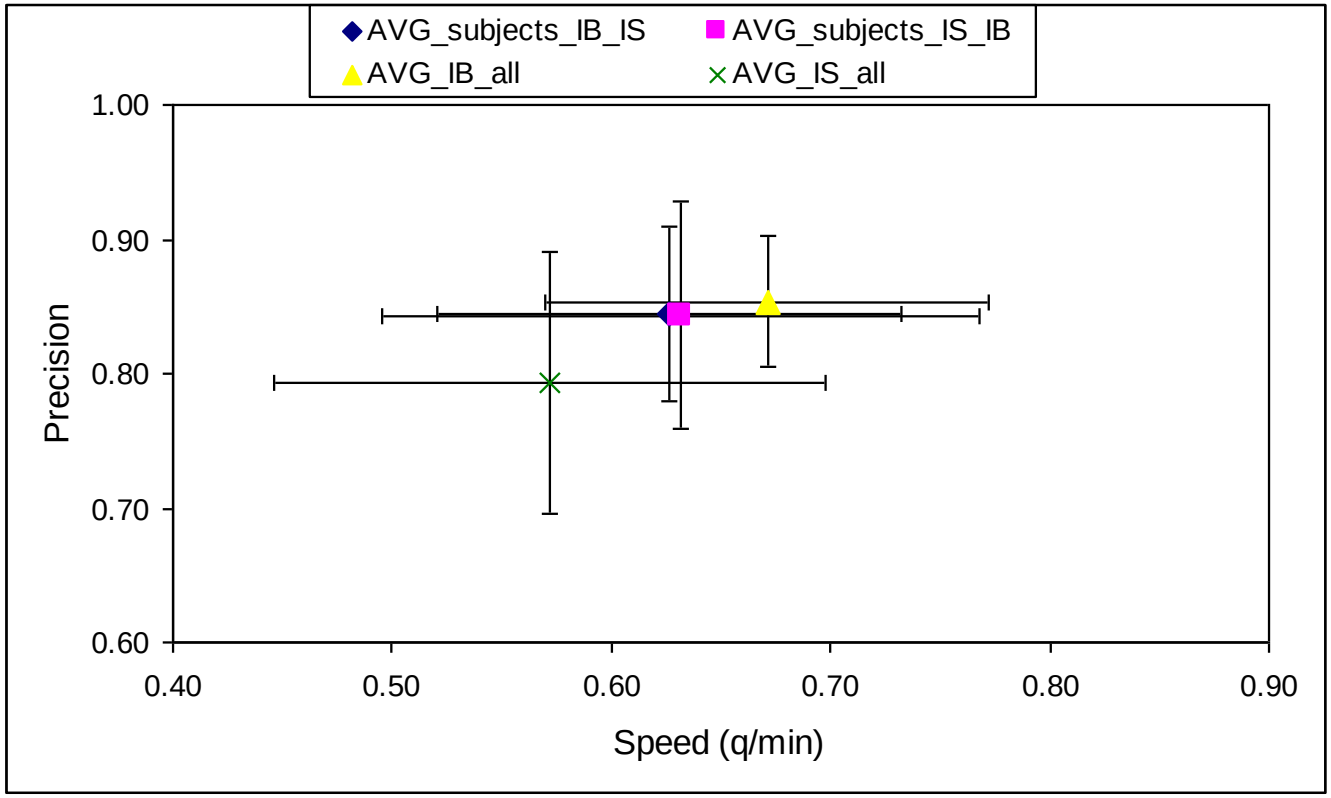
- IS1008c – Remote Control Design
  - According to the manufacturers, the casing has to be made out of rubber. /// According to the manufacturers, the casing has to be made out of wood.
  - Christine suggested that customers might want to submit their own design via the internet as custom orders. /// Christine suggested that customers would not be interested in custom design and prefer off-the-shelf products.

- See also the practical session

Results of applying the BET to the TQB browser

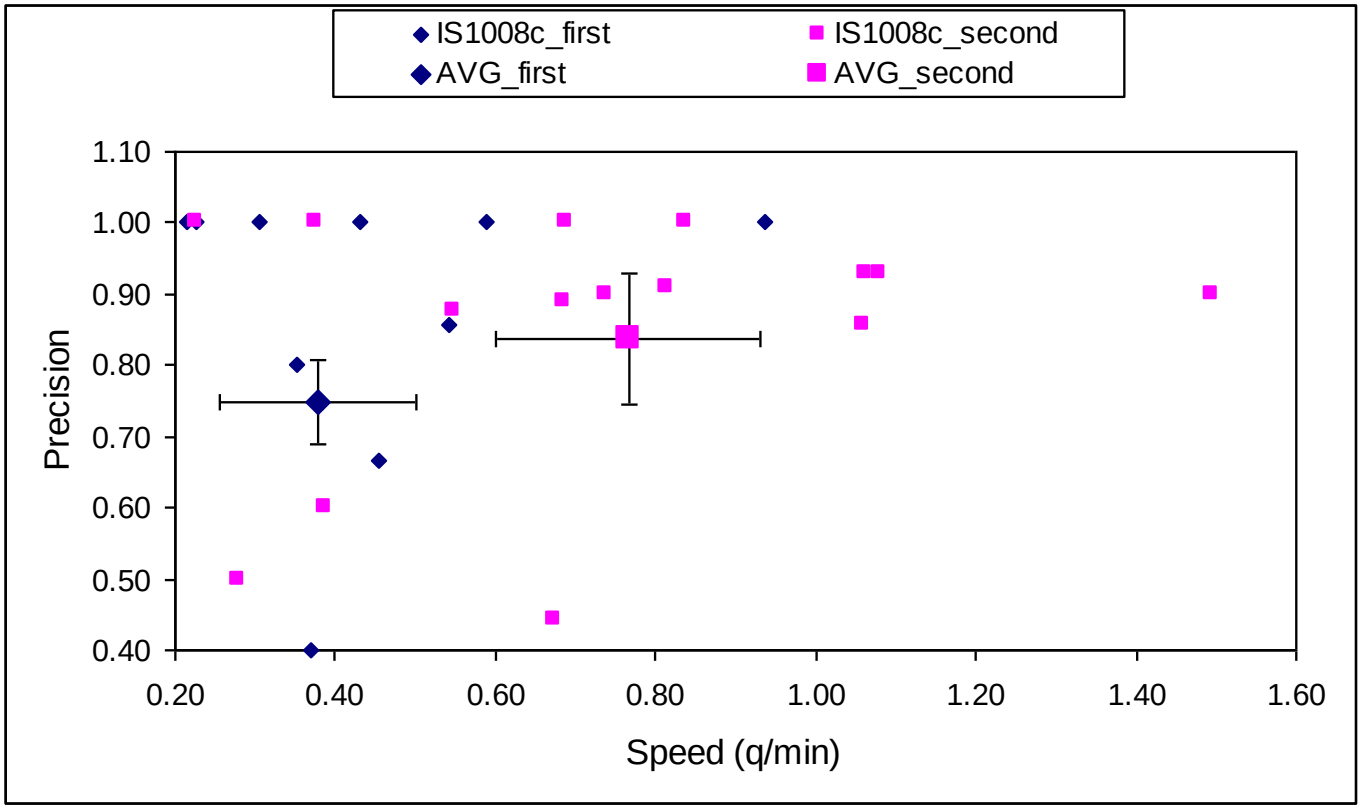
- 28 students (in translation, no experience with m.b.)
- half started with IB4010 and continued with IS1008c (IB_IS)
- the other half did the reverse order (IS_IB)
- time: about 25 min. for IB4010 and about 13 for IS1008c

Average TQB speed and precision



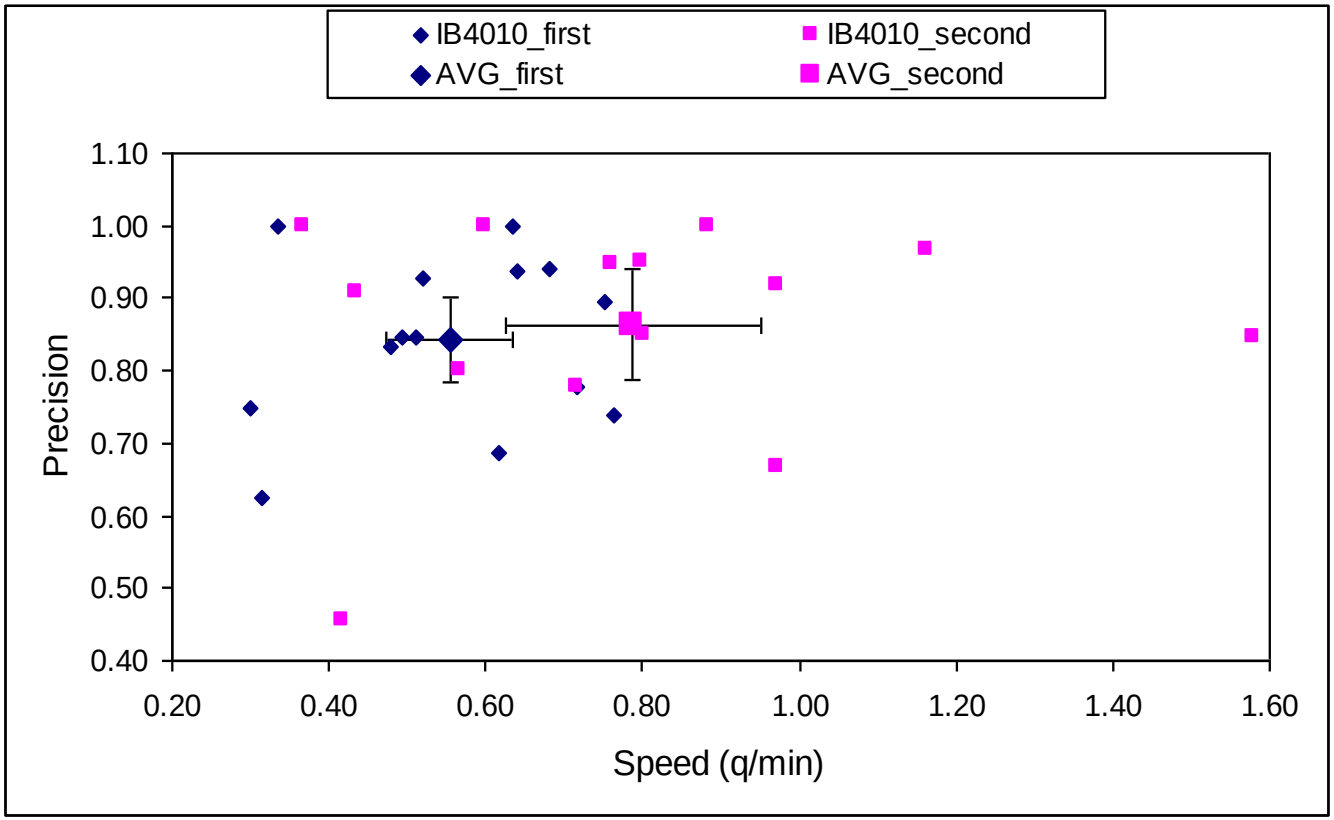
- Is performance across groups similar? Yes
- Are the questions over the 2 meetings of comparable difficulty?
  - almost, but IB4010 seems easier than IS1008c, though it's longer

IS1008c: Individual scores and averages when it is seen first (blue diamonds) *vs*. when it is seen second (pink squares)



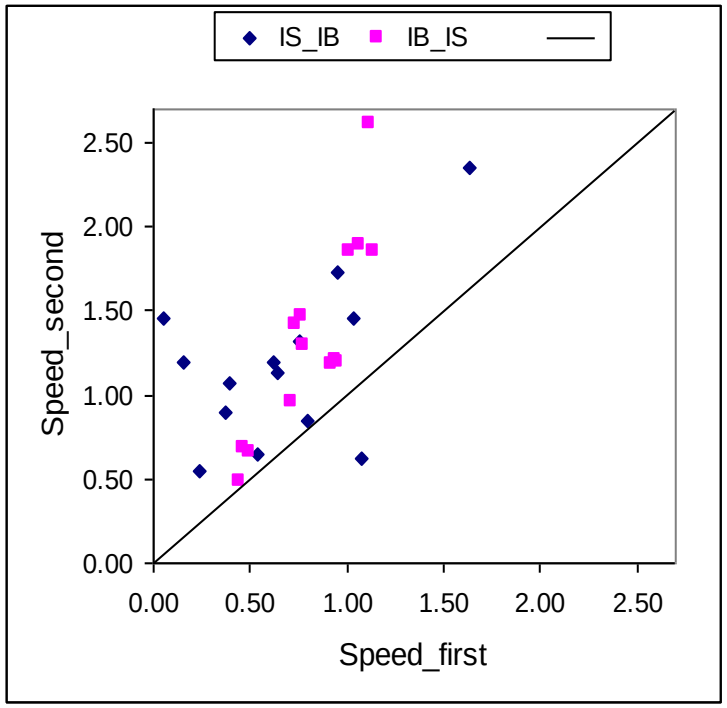
- Speed increases when IS1008c is seen second
- Precision does not increase significantly

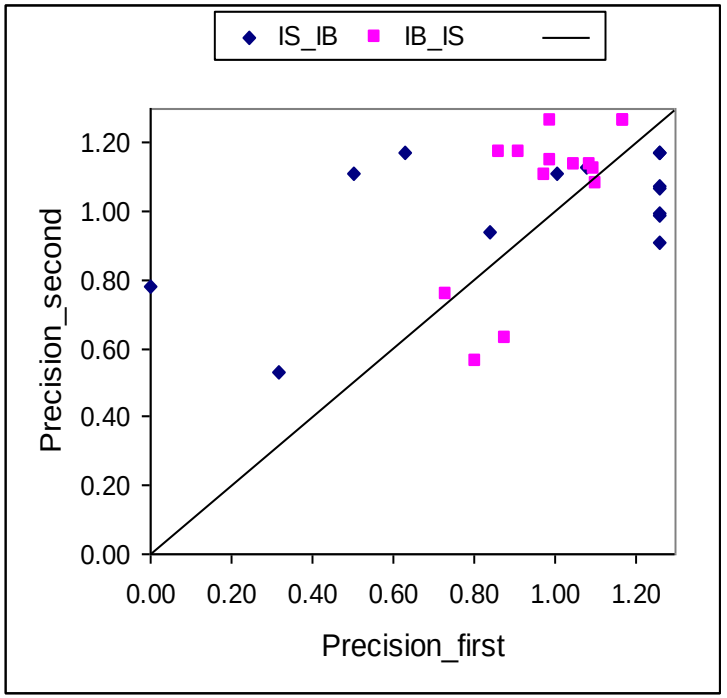
IB4010: Individual scores and averages when it is seen first (blue diamonds) *vs*. when it is seen second (pink squares)



- (results are comparable to IS1008c)

A view of the training effect (1st vs. 2nd meeting): speed improves, but precision not much

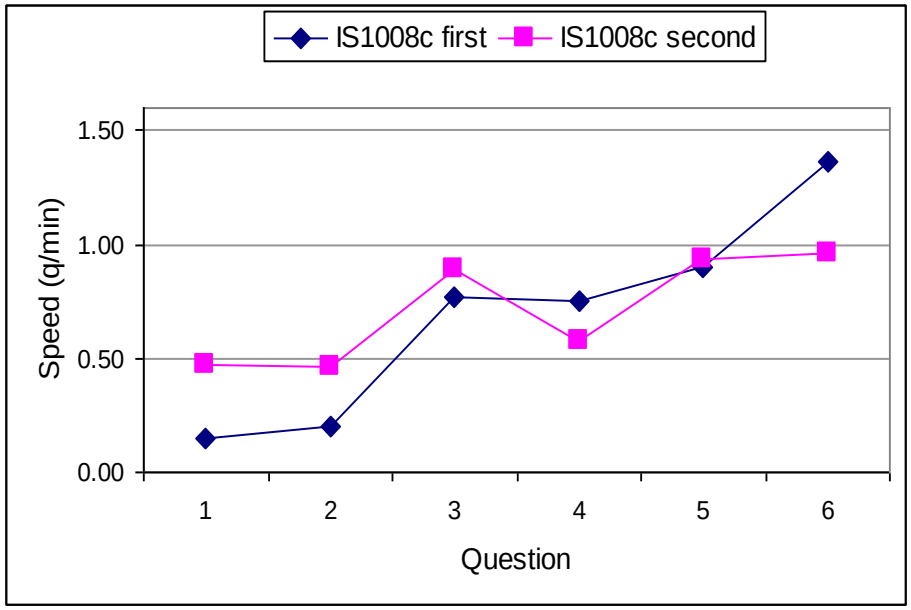


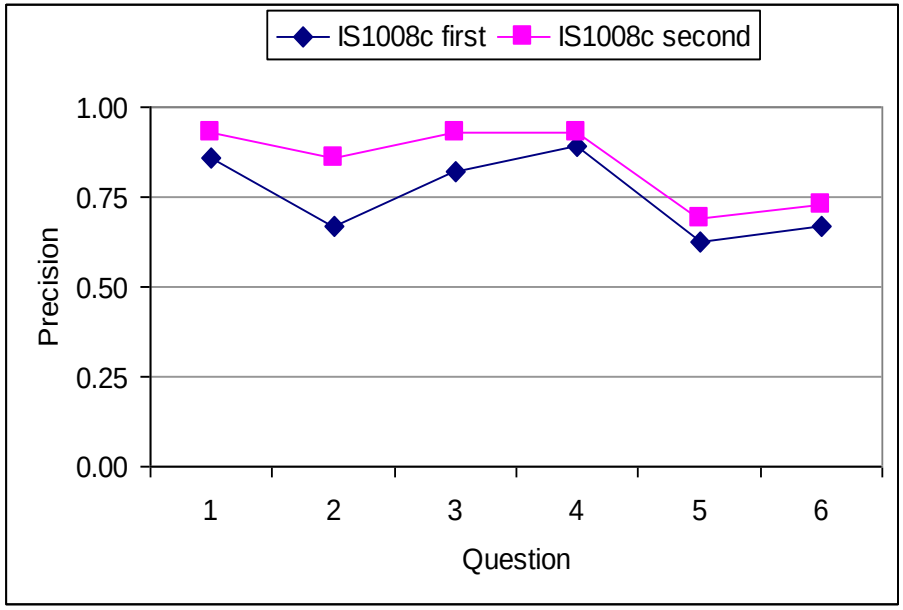


- Here, values for each meeting are normalized by the overall average for the meeting to compensate for variations in difficulty

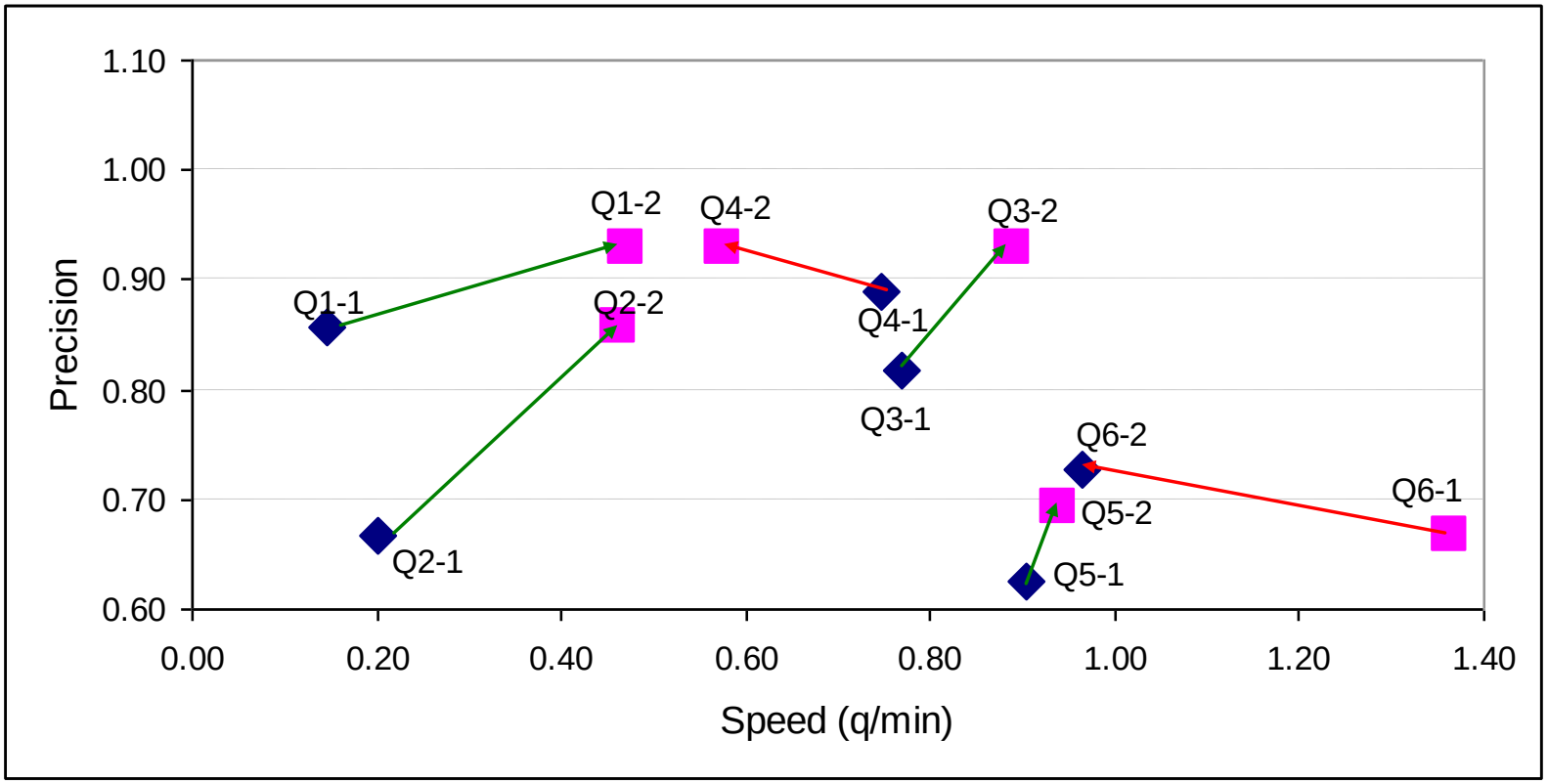
Speed and precision per question: IS1008c

group **IS**_IB (diamonds), group IB_**IS** (squares), first 6 questions



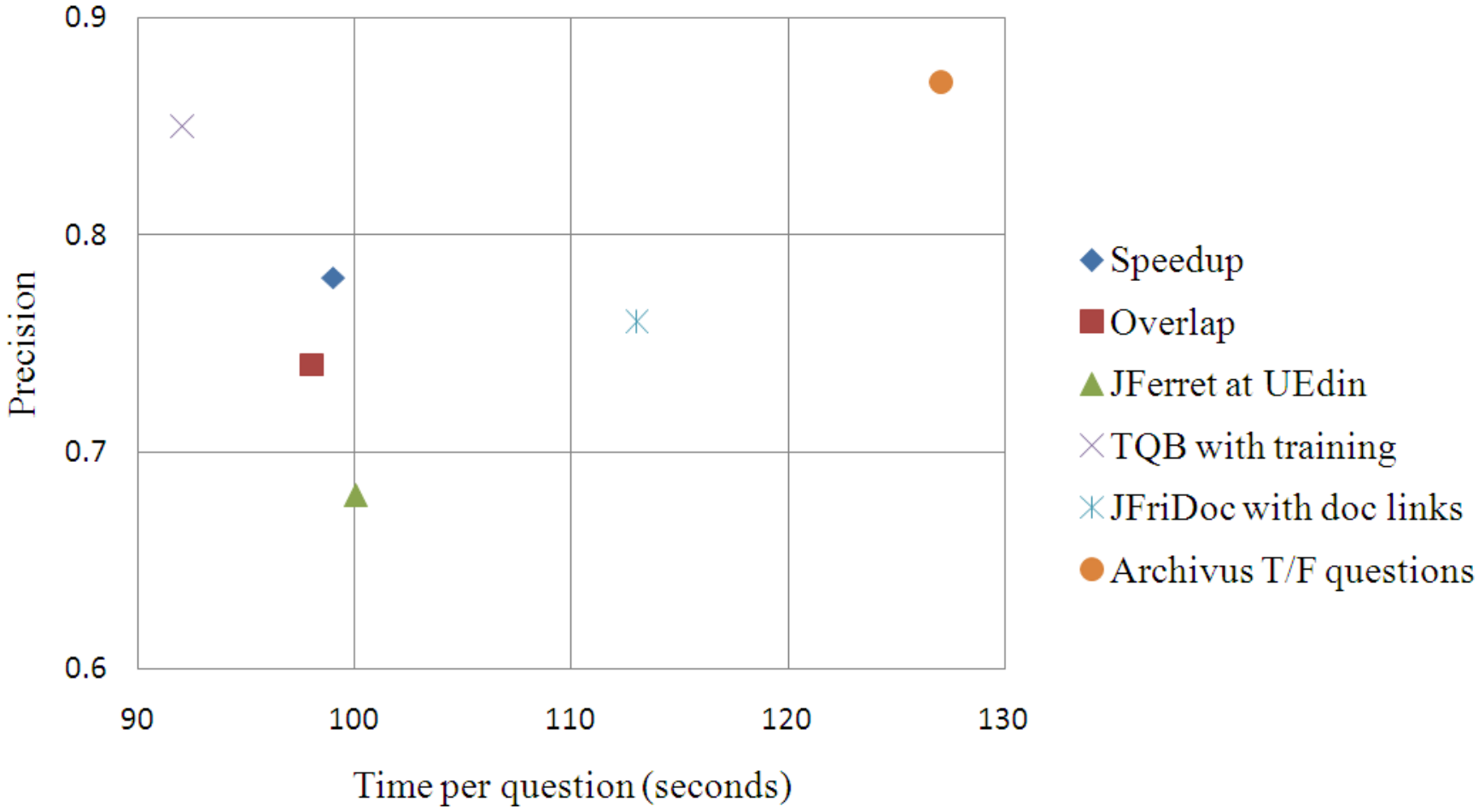


IS1008c: precision for first 6 questions, when the meeting is seen first *vs*. when it is seen second



- Green arrows: precision and speed increase
- Red arrows: precision increases but speed decreases

Sample BET results for several browsers



Sample BET results: nb. of subjects (NS), average time per question (T), precision (P), with confidence intervals (±CI)

Browser	Condition	NS	T(s)	CI	P	CI
Audio-based	Speedup	12	99	26	0.78	0.06
browsers [13]	Overlap	15	88	23	0.73	0.08
JFerret	BET set (pilot)	10	100	*43*	0.68	*0.22*
[20]	5 gisting questions	5	<180	0	0.45	*0.34*
[21, p. 210]	5 factual questions	5	<180	0	0.76	*0.25*
TQB	1st meeting	28	228	129	0.80	0.09
[13]	2nd meeting	28	92	16	0.85	0.06
	Both meetings	28	160	66	0.82	0.06
FriDoc	With speech /	8	113	n/a	0.76	n/a
[16]	document links					
	Without links	8	136	n/a	0.66	n/a
Archivus	T/F questions	80	127	*36*	0.87	*0.12*
[11, Ch. 6.6]	Open questions	80	n/a	n/a	0.65	*0.22*
AutoBET	Movie club meeting	5fCV	<1	n/a	0.57	0.06
[8]	Remote control meeting	5fCV	<1	n/a	0.64	0.18

Conclusions: lessons learned

- Requirements depend on how subjects are questioned
  - a fixed specification cannot be set from the start
  - user-studies must be gradually focused toward a tractable task
- Technology providers have various views of what is “useful”
  - they tend to evaluate technology from their own perspective
  - their view of HLT utility might differ from users’ view
- Combine user-driven and technology-driven approaches
  - go back-and-forth from the users’ perspective to the developers’ one
  - specify a reasonable task and the related evaluation method

→ here, the fact-finding task and the Browser Evaluation Test

Future of meeting browsers

- Some existing products
  - conference browsers: Klewel (Idiap), SMAC (CERN)
  - *potential commercial success*
- Extension #1: automatic browsers
  - directly answer questions from users
  - our practical exercise: discriminate BET pairs automatically
  - *spoken QA during conversations*
- Extension #2: query-free automatic browsers
  - answer implicit queries for accessing meeting archives
  - *context-sensitive just-in-time information retrieval*

References

- A. Popescu-Belis, D. Lalanne, and H. Bourlard, “Finding Information in Multimedia Meeting Records”, *IEEE Multimedia*, vol. 19, p. 48-57, 2012.
- P. Wellner et al., ‘‘A Meeting Browser Evaluation Test,‘‘ *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (CHI 2005)*, ACM Press, 2005, pp. 2021-2024.
- A. Popescu-Belis et al., ‘‘Towards an Objective Test for Meeting Browsers: The BET4TQB Pilot Experiment,‘‘ *Proc. 4th Workshop Machine Learning for Multimodal Interaction (MLMI 2007)*, LNCS 4892, Springer-Verlag, 2008, pp. 108-119.
- S. Renals et al., *Multimodal Signal Processing: Human Interactions in Meetings*, Cambridge Univ. Press, 2012.