Question answering over multimedia meeting recordings: assistance or automation?

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Information access within meeting recordings can be assisted by meeting browsers, or can be fully automated in a question-answering (QA) approach. To evaluate performance, several groups of users have been asked to discriminate true vs. false parallel statements about facts in meetings, using several browsers. A review of the results indicates that state-of-the-art browsing speed is 1.5–2 minutes per question, and precision is 70%–80% (vs. 50% random guess). An automatic QA algorithm is introduced for the same task, using passage retrieval over a meeting transcript, followed by statement discrimination. The algorithm scores 59% accuracy for passage retrieval (vs. $< 1\%$ random guess), but only 60% on combined retrieval and discrimination, while largely outperforming humans for speed, with less than 1 s per question. The degradation on ASR compared to manual transcripts is acceptable, but scores decrease quickly when shorter and shorter automatic summaries are used. Automatic QA is a promising enhancement to meeting browsers, which helps to direct human users to relevant information in meetings.

1 Introduction

The increasing amount of multimedia recordings, in particular of human meetings, raises the challenge of accessing the information contained within such recordings. The automatic processing of language and other modalities from meeting recordings involves a large variety of component technologies such as speech recognition, diarization, summarization, but also document and video processing. The use of such technologies has often been justified as facilitating the access to the information contained in meeting recordings, by transforming raw data into more and more abstract layers of representation.

Even if the output of speech, language and multimodal processing technology is in many cases not directly usable by humans for an information access task, this output is either rendered via meeting browsers or used as input for more abstract processing modules. For instance, automatic summarization of spoken language in meetings can use the output of speech recognizers along with utterance segmentation and dialogue act recognition. The applicative objectives of meeting processing techniques naturally raise the question of their actual usefulness in the intended context, namely for a general-purpose information access task.

In this paper, a question answering (QA) approach is adopted for the evaluation of tools that use meeting processing in order to enhance access to meeting recordings. The focus is on the access to specific bits of information from the meeting, as opposed to tasks that require an abstraction over the entire meeting. The main goal is to present several evaluations of meeting browsers that were carried out with similar resources and metrics, and to compare the performance of humans using meeting browsers on a QA task with an automatic QA system designed for this task. 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. Furthermore, the results demonstrate the utility of automatic QA techniques as an assistant tool for meeting browsers.

The paper is organized as follows. Section **??** introduces the data and the meeting browsers under study. Section **??** describes the construction of the evaluation resources and the overall evaluation protocol named the BET. Section **??** puts together the results obtained when several browsers –

audio-based, transcript-based, or document-based – were evaluated using the BET, in terms of subjects' speed and precision to answer questions about a meeting. Section **??** describes an automatic QA system designed to take the BET, in two stages: passage identification and question disambiguation. Section **??** provides the score of the QA system in several conditions – on manual transcripts, on ASR, on automatic summaries – and compares them with those of human subjects using the browsers. This comparison shows that humans outperform the system in terms of precision, but are by far slower, and that the best use of the QA system would be as an assistant for relevant passage identification.

2 Meeting browsers: data and technology

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 (**?**; **?**; **?**), 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 as well as other potential end-users to access the information enclosed in the recordings.

Many scenarios in which meeting browsers answer specific user needs have been described (**?**; **?**; **?**; **?**), although more user-centric studies of meeting technology are still needed. An important distinction has been made between two types of functions, which can both be required depending on the intended use of the browser: "gisting" is the synthesis of the essential[1] information contained in meeting recordings, akin to summarization and including decision, action points, etc., while "information access" targets precise facts located in specific sections of a meeting. 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, though quite often the speech transcript plays a central role.

In this paper, we focus on the information access task over a given meeting, in which 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 task (**?**; **?**), using different types of features extracted from multimedia recordings: speech, transcript, annotations, documents, and videos. The issue of evaluating these browsers is discussed in the next section; here, we outline the functionalities of several browsers that were submitted to evaluation using a common protocol, with results that were available to the authors.

All of them are summarized in (**?**), in fact, except the Idiap audio-based.

JFerret: (**?**; **?**)
Idiap Audio-based: (**?**; **?**)
TQB: (**?**; **?**)
FriDoc: (**?**)
Archivus (from Mirek's thesis): (**?**; **?**)

3 Evaluation of meeting browsers and the BET

3.1 Overview of existing methods

Some landmarks in the evaluation of interactive software, especially multi-modal dialogue systems, which is still an open problem (**?**; **?**; **?**). As the task of meeting browsing does not impose specific functionality requirements, the most appropriate technique is task-based evaluation, or evaluation in use. The main parameters to be evaluated, in this approach, are *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 (**?**), 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 (**?**) but these must be adapted to the architecture of each system, and do not depend only on the task.

Inspiration from TREC QA task which started in 1999 (**?**; **?**). Mention the TREC procedure to get questions: questions were submitted by participants along with assessors, FAQFinder, orgaizers (result: 1337 q, of which 200 were selected). At TREC 2003, the test set of questions contained 413 questions (of 3 types: factoid, list, definition)

[1] 'Essential' or 'important' is always relative to a specific point of view or task.

drawn from AOL and MSNSearch logs. (**?**). Trec 2003 had an "exact answers" track, but also a "passage" track, where it was enough to return the passage – systems designed for this track share inspiration with the automatic system described below (Section **??**)..

An evaluation task for interactive question answering was proposed in iCLEF, the Interactive track for the Cross-Language Evaluation Forum (**?**), 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.

Other approaches to meeting browser evaluation: quote from my chapter in AMIDA book the TBET (**?**) and the "audit-based" approach for summary evaluation (**?**). Explai how it works.

In (**?**), they also evaluate speech compression by finding out whether the compressed version allows as good a ranking of utterances (according to importance) as without compression. But very specific conditions!! Is this a realistic task? (No.)

3.2 The Browser Evaluation Test

The Browser Evaluation Test, or BET (**?**; **?**). (QA framework)

Overview. Recall that this is for information access, which does not cover all the uses of a meeting browser.

How the questions were collected (unlike TREC, we don't ask participants!), examples, i-a agreement.

Three meetings from the AMI Corpus (**?**) were selected for the observation collection procedure: IB4010, IS1008c, and ISSCO-Meeting_024. The meetings are in English, and involve four participants, native or non-native English speakers. In the first meeting, the managers of a movie club select the next movie to show; in the second one, a team discusses the design of a remote control; in the third one, a team discusses project management issues. Although the first two meetings are in reality enacted by researchers or students, the movie club meeting (IB4010) appears to be more natural than the remote control meeting (IS1008c), probably due to the familiarity of the participants with the topic. For each of these three meetings, BET observations were collected, edited and ordered, this resource being now publicly available at `http://mmm.idiap.ch`. In the evaluations below, the order based on importance was kept constant.

For these meetings, respectively 222, 133 and 217 raw observations were collected, from respectively 9, 6 and 6 observers, resulting in respectively 129, 58 and 158 final pairs of true/false observations. As initial observations are grouped according to their similarity, as explained above, the average size of the groups (1.72, 2.29 and 1.37 observations per group) provides a measure of inter-observer agreement. While these values are not very high with respect to the number of observers, it is more eloquent to consider only the agreement for the observations that were answered by at least half of the subjects in the experiments on TQB (i.e. 16 for IB4010 and 8 for IS1008c). As these were ranked by importance, the average number of observers having made these observations was around 5 for both meetings, i.e. 55% and 83% of the observers agreed upon them.

Examples of the five best observations are in Table **??**. The BET questions are available upon request from `http://mmm.idiap.ch` – aren't they?!

Typically how it is applied. Duration, training, comparison.

What scores are measured (+ ref to effectiveness, efficiency, user-satisfaction (in some cases questionnaires were collected)) and computed.

The duration allowed for each meeting was half the duration of the meeting: 24'40" for IB4010, and 12'53" for IS1008c; the timing was managed by the BET master interface.

It is crucial to discuss how exactly the speed is computed! Normally, we should focus on times (which are an additive grandeur) and not on speeds. So, average time per person, then per group, etc. And not speeds!!!!! In previous papers (**?**; **?**) they (!) got the figures wrong for the audio-based browsers because they were averging speeds:

Let us note subjects with $s = 1 \ldots S$, questions answered by subject s with $q = 1 \ldots Q(s)$. Time to answer question q by subject s is noted $t(q, s)$. Average time per question for each subject is $\bar{t}(s) = T_0/Q(s)$ where T_0 is the fixed time allowed for the experiment. If not fixed (e.g. because we average over two meetings!):

	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 1: First five most quoted observations of interests, in true/false representation.

$$\bar{t}(s) = \frac{1}{Q(s)} \sum_{q=1}^{Q(s)} t(q,s)$$

Average time per group should be:

$$\bar{T} = \frac{1}{S} \sum_{s=1}^{S} \bar{t}(s) = \frac{T_0}{S} \sum_{s=1}^{S} \frac{1}{Q(s)}$$

and not:

$$\bar{T}' = \frac{S \cdot T_0}{\sum_{s=1}^{S} Q(s)}$$

Note also that average of speeds (not a very good idea) is not directly linked to the speed computed using average time (which should be a slighlty better idea)?????????:

$$\frac{1}{S} \cdot \sum_{s=1}^{S} \frac{Q(s)}{\sum_{q=1}^{Q(s)} t(q,s)} \neq \frac{\sum_{s=1}^{S} Q(s)}{\sum_{s=1}^{S} \sum_{q=1}^{Q(s)} t(q,s)}$$

4 Overview of BET experiments with meeting browsers

The evaluations of browsers: we already briefly outlined each browser in Section **??**, so here we give results in text, then in the synthetic Table comparing all results, in terms of precision and average time, Table **??**.

Ferret or JFerret. trial run in the original BET paper: Trial Run tests In the trial run, we tested a total of eleven women and thirteen men primarily from academia, whose average age was 35. All subjects were given 22 minutes to answer questions about the 44-minute trial run recording. 3 conditions; player, base, Ferret. Meeting was issco-024 on furnishing. 10 subjects in F1; results 0.60 +- 0.26 q/min, and prec 67.7% +- 22.4

And some results in Simon's/Steve's paper: In (**?**), they report an evaluation using 10 tasks derived from BET questions. They do not have closed form answers, e.g. "what is the status of the project?" We have JFerret with a slightly different evaluation in (**?**, p. 210-211): They were given the 10 tasks above in random order. They were asked to answer them as accurately as they could, using the browser. We imposed a time limit of 30 min for the whole experiment to stop subjects from simply playing the recording from beginning to end to answer each question. 5 question rather about "gisting", and 5 about facts, very close to BET... – again with open form answers.

Audio-based browsers. Idiap's/Sheffield experiments with: "player", and 3 conditions (quote paper – I think you can also quote the deliverable, (**?**), but check it's available online and that URL appears in refs).

In another series of experiments (**?**) ???, conducted by the IDIAP Research Institute and the University of Sheffield, four meeting browsers or "conditions" were tested with the BET, in a slightly different setting than the one described above. Usable data was obtained from 39 subjects: each subject performed a calibration task (answering questions using a very simple browser), and one of the following browsers: base (15 subjects), speedup (12 subjects), and overlap (12 subjects). Unlike TQB, none of these meeting browsers relied on manual annotation of the data or on human transcripts. The ISSCO-Meeting_024 was used for

Browser	Condition	#Subjects	Time/q (s)	Stdev*	Precision	Stdev*
Player	??					
Audio-based browsers	Speedup	12	99	26*	0.78	0.06*
	Overlap	15	88	23*	0.73	0.08*
JFerret-style	Gisting (5 q.)	5	180***	0	0.45	0.34
	Factual (5 q.)	5	180***	0	0.76	0.25
	BET set (pilot)	10	100	43	0.68	0.22
Transcript-based (TQB)	1^{st} meeting	28	228	129*	0.80	0.09*
	2^{nd} meeting	28	92	16*	0.85	0.06*
Document-based (FriDoc)	With links	8	113	n/a	0.76	n/a
	Without links	8	136	n/a	0.66	n/a
Archivus** multimodal	T/F q	80	127	36	0.87	0.12
	Open q.	80	==	==	0.65	0.22

Table 2: Comparative results of several meeting browsers. Average time needed by subjects to answer a question, and average precision. STDs (or CIs at 95% when marked with a *) are in absolute values (of precision or time). Put numbered notes to explain many differences.

calibration, and the other two meetings (IB4010 and IS1008c) were used alternatively in the different conditions.

The calibration condition presented a large slide view, 5 video views, the audio, a timeline, and slide thumbnails. The base condition played audio and included a timeline, scrollable speaker segmentations, a scrollable slide tray, and headshots with no live video. The speedup condition was exactly like the base condition except that it allowed accelerated playback with a user-controlled speed between 1.5 and 3 times normal speed. The overlap condition duplicated the speedup condition by offering simultaneously the first half of meeting on the left audio channel of the subject's headphone, and the second half of the meeting on the right channel, requiring the subjects to focus on one channel at the time. Raw performance scores for both meetings were as follows for the three conditions (see (**?**, Section 3) for more details). For the base condition, average precision and speed were respectively 0.77 and 1.2 questions per minute; for the speedup condition, 0.83 and 0.9 questions per minute; and for the overlap condition, 0.74 and 1.0 questions per minute. The average precision is generally below the values obtained by TQB (0.84 ± 0.05 for TQB), while speed is always higher (0.63 ± 0.09 for TQB). These results are quite surprising, as TQB provides access to the transcript, which should considerably improve its information extraction capabilities. In addition, although TQB subjects were not native English speakers unlike those of the other two browsers, data from TQB shows that proficiency is in fact better correlated with precision (at 0.65 level) and much less with speed, therefore the proficiency factor might not explain the difference in precision. Other factors must thus be found, by analyzing experimental logs, to account for these differences.

If we remove the two crazy guys from Overlap, time is 98 +-22 and 0.74 +- 0.10.

BET4TQB

A synthesis of the results presented elsewhere (**?**; **?**).

TQB was tested with 28 subjects, students at the University of Geneva, mainly from the School of Translation and Interpreting. Results from 4 other students were discarded for not completing the two meetings. The average proficiency on a 4-point scale (from 'beginner' to 'native') was 2.6, median value being 3 ('advanced'). Half of the subjects started with IB4010 and continued with IS1008c, and the other half did the reverse order, thus allowing for differentiated results depending on whether a meeting was seen first or second within the trial.

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 time needed is 160.1 ± 66.3 seconds per question. 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: time needed and precision are 104.0 ± 15.16 and respectively 0.85 ± 0.05 for IB4010, both higher (!) than the respective values for IS1008c, 216.1 ± 128.4 and 0.79 ± 0.10. (Well, but at 90% it would work!) If the statistical significance was higher, one could conclude that IB4010 is easier than IS1008c from the BET perspective. The other values are in Table**??**.

JFriDoc

The only reference describing the browser is (**?**), however, the results are personal communication. Compared enabled vs. disabled document-centric browsing, i.e. with vs. without temporal links on documents 8 students tested both options, with vs. without, on different meetings had to answer 12 questions each, in maximum 3 minutes each

Archivus

An evaluation of Archivus is available in Mirek's thesis): (**?**)

Says Mirek: these results cannot be directly compared with the results obtained in BET evaluations for other browsers (audio-based browsers and TQB), mainly because we have evaluated the browser on a different set of tasks (compared to BET) and we have six-times larger search database!!.

For our work, we only use the data from the second twenty minute experimental sessions, during which all modalities were available to the users (with the exception of pen and mouse, which were never available simultaneously), and keeping 80 users to balance conditions (of training, but also of test: half had mouse-voice-keyboard and half pen-voice-keyboard). (For training, they had a limited number of modalities).

The tasks included true-false questions (statements) like "The budget for the room furnishing was 1000CHF" and short-answer questions like "Who attended all meetings?" (so BET-style, but not exactly BET) The data ... Overall, our database includes 6 meetings (192 minutes of video data) held in English by a total of 8 different participants with typically 4 participants in one meeting. The rooting scenario of four meetings is a room furnishing, while the other scenario is a movie club meeting and a meeting to determine a design of a remote control.

He evaluated also on "short answer" questions, which we also report here, and they have a different baseline.

Duration per task. When includes system response times 127s (+-36s), but excluding system response times 91s (+-31s). So the WOZ system requires 36 seconds to answer!! I'd rathergive the figures for "include" because that's true for the others as well. But the system takes on average 36 sec to reply, due to the Wizard of Oz. In fact, figures for time are given jointly for the T/F and open questions, so I cannot but report them as they appear.

5 Automatic BET Question Answering

Automatic BET: explain the processing and how it works.

Works on transcript (tested with both human and ASR).

Describe the stages, the pre-processing, also here the "training".

There are four main steps in this stage. 1) Removing characters not related to words as punctuations: comma, dot, quotation mark, semicolon or exclamation, etc. 2) Converting abbreviated words into full forms: We've becomes We have, I'll becomes I will, etc. 3) Converting numeric forms to text forms: 34 becomes thirty four, 2nd become second, etc. 4)Removing stopwords: For the reason that the questions are indirect-speech statements while the meeting transcript is a direct-speech report, we have to remove words related pro- nouns: personal pronouns, possessive pronouns, demonstrative pronouns as a list of stop words. Some prepositions, conjunctions and articles (a, an, the) are removed because they offer little sematic information and only to reduce cost of computing. Then, both the questions and the transcript are splitted into sequences of words. Stemm is used in order to remove in inflectional affixes. This may increase the signal from contentful words. We use the stemmer programme of Porter [8]. Lexical meaning is also extended by synonyms and lemmas using WordNet [7]. All words are transformed into lower-case forms. Hence, there are not proper noun or pronoun any more. They will be treated in the same way. Each question word is considered as a record of 3 fields: Original word, Lemma of word and Stem of Lemma. In fact, lemmatization sometimes returns more than one lemmatized word for original word, hence, this field is established as an array of strings. Table 4.1: Word splitting for question Position Word Lemma Stem 1 Mirek Mirek Mirek

2 had have have 3 not not not 4 received received receive receiv ... For transcript, each word record has 6 fields in which synonyms are narrowed by part of speech (PoS) produced by QTAG [?]. The field “Speaker” indicates name of person who used this word. This field is important to questions that demand a verification “who did what” - a frequent type of questions in experimental data. Position is the original position of word before removing stopwords. This field is used to calculate distance among matched words afterward in the phase 2, true-false question answering. All data are processed before running main algorithm and stored in RAM in order to speed execution of the programme.

“synonyms matching” in order to bring the highest score for the current passage. - The frequency of one matched word is not used to increase the score. However, in the case “multiple matching”, if one word repeats more than one time in both question and 17 transcript, the score will be calculated as the minimum number of appearance of this word between the question and the transcript.

For discrimination:

Require: Passage1, Passage2 score1 = Score of Passage1 Score2 = Score of Passaag2 if score1 ¿ score2 then return 1 else fscore1 ? score2g d1 = Distance among matched words between passage1 and question1 d2 = Distance among matched words between passage2 and question2 if d1 ¡ d2 then return 1 else fd1 ? d2g i = 2; while Score1 ? Score2 do score1 = number of matched i-gram search between passage1 and question1 score2 = number of matched i-gram search between passage2 and question2 end while if score1 ¿ score2 then return 1; else return 2; end if end if end if

6 Results for Auto BET and comparison

6.1 How to evaluate the 2 phases

In order to assess the correctness of a retrieved passage, a reference passage made by hand is used to compare with it. The size of this reference passage is reduced as small as possible but it still contains essential words for answering question. For that reason, even some keywords as name of topic being discussed is not necessary to be included in the reference passage. The information of the reference passage is the position of its first word and its last word in current transcript. All transcript words are numbered as their position in the transcript by passing name of speakers

If candidate passage and reference passage are overlapped each other, candidate passage is considered correct. The number of overlapped words is fixed as one words or multiple words. If the system is used to help users locate position of answer information, one overlapped will be accepted as well.

In order to evaluate the performance of the system, we want to divided questions into two classes: straightforward and deductive questions. But as being addressed in the chapter 3, it is difcult. According to our subjective analysis, we find 42 (36.21%) questions for IB4010 that requires a deduction to answer and only 4(8%) such questions for IS1008c. In other words, we have 63.79% of easy IB4010 questions and 92% of easy IS1008c questions.

So, IB4010 = Maximal score by subjective analysis 63,79% ((in both cases?) For IS1008c, maximal score by subjective analysis is 92% (in both cases?).

What is a baseline score for AutoBET? In which, score of random method is calculated as below: Total passages = [(transcript size - window size)/window step] + 1 Correct passage = 2*(window size / window step) Random score = correct passage/(total passages)) IB4010 score = (2*8*7/7) / ([(4872 - 8*7)/7) = 0.33) IS1008c score = (2*8*7/7) / ([(2059 - 8*7)/7) = 0.78

6.2 The data

Or maybe not much to say? We have 116 pairs of true-false analogous questions for IB4010 and 50 pairs for IS1008c. – because we used only those that were given to humans after cleaning.

6.3 The results

Give here the scores for the AutoBET: best scores, results on ASR, and on summarization.

Compare these scores to humans, discuss draw conclusions (or below?): it is a useful tool for locating answers, not necessarily finding it. For finding it, work on entailment should be used (quote some broad references). Or maybe this goes into the general conclusion? Future work: need to use semantics, too.

Optimization of scores: brute force search for windows size and step, that is, With the purpose of calculating the average of highest scores, the algorithm repeates on training data for different parameters from (search window size = 1 x L, search

	Passage retrieval				**Disambiguation**			
	IB4010		**IS1008c**		**IB4010**		**IS1008c**	
Condition	**Acc.**	**Stdev**	**Acc.**	**Stdev**	**Acc.**	**Stdev**	**Acc.**	**Stdev**
Random	.0033	n/a	.0078	n/a	0.50	n/a	0.50	n/a
Unigram matching	0.27	0.15	0.54	0.21	0.37	0.14	0.36	0.21
N-gram matching	0.32	0.15	0.50	0.19	0.43	0.17	0.42	0.11
N-gram & speaker weighing	0.55	0.14	0.62	0.16	0.57	0.06	0.64	0.18

Table 3: Passage retrieval accuracy for the two meetings. Stdev is computed using 5-fold cross validation.

window step = 1 x L) to (search window size = 13 x L, search window step = 13 x L) in which L is the size of question. The most suitable parameters for IB4010 is the pair (10,3) and (4,1) for IS1008c, and those are used in the paper.

Results for passage retrieval given in Table **??**. Remind maximal scores by "subjective" analysis. And combined results in Table **??** too.

And now, a comparison with human scores, in Table for IB4010 **??** and in Table for IS1008c **??**.

What would be nice is to have also some results on passage retrieval over summaries (disambiguation doesn't work well). Compare degradation over a "reference" summary (?), over Gabe's one, and over a random excerpt reduced at the same size.

7 Conclusion

Probably just a recap of the whole paper rather than new ideas.

Acknowledgments

amida trainee for qal – im2 support for apb.

IS1008c	**As first meeting**		**As second meeting**		**System**	
Question	**Average**	**Average**	**Average**	**Average**	**Passage**	**Combined**
number	**time (s)**	**precision**	**time (s)**	**precision**	**accuracy**	**accuracy**
1	303	0.93	143	0.71	0	0
2	105	0.93	66	1.00	1	1
3	118	0.71	89	1.00	1	1
4	207	0.86	206	0.86	1	1
5	65	1.00	37	0.93	0	1
6	58	0.93	53	1.00	1	1
7	61	0.93	52	0.71	1	1
8	129	0.71	85	0.79	1	1
Average	303	0.88	92	0.88	0.75	0.88

Table 4: For IB4010 Comparison of system with humans over the questions answered by all subjects (14). Of course, system speed is very high (after training), less than 1 s per question.

IS1008c	**As first meeting**		**As second meeting**		**System**	
Question	**Average**	**Average**	**Average**	**Average**	**Passage**	**Combined**
number	**time (s)**	**precision**	**time (s)**	**precision**	**accuracy**	**accuracy**
1	410	0.86	127	0.93	1	1
2	299	0.67	129	0.86	1	1
3	78	0.82	67	0.93	1	1
4	80	0.89	104	0.93	1	1
5	66	0.63	64	0.69	1	0
6	44	0.67	62	0.73	0	0
7	24	1.00	48	0.82	1	0
8	66	0.67	94	0.64	0	1
Average	133	0.77	87	0.81	0.75	0.63

Table 5: For IS1008c: Comparison of system with humans over the questions answered by several subjects (from 14 to 3 as first meeting, from 14 to 11 as 2nd meeting). Of course, system speed is very high (after training), less than 1 s per question.