IDIAP RESEARCH REPORT

AUTOMATIC VS. HUMAN QUESTION ANSWERING OVER MULTIMEDIA MEETING RECORDINGS

Quoc Anh Le Andrei Popescu-Belis

Idiap-RR-13-2009

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

Automatic vs. human question answering over multimedia meeting recordings

Quoc Anh Le[1,*], *Andrei Popescu-Belis*[2]

[1]University of Namur, Belgium
[2]Idiap Research Institute, Martigny, Switzerland
lequocanh@student.fundp.ac.be, andrei.popescu-belis@idiap.ch

Abstract

Information access in meeting recordings can be assisted by meeting browsers, or can be fully automated following a question-answering (QA) approach. An information access task is defined, aiming at discriminating true vs. false parallel statements about facts in meetings. An automatic QA algorithm is applied to this task, using passage retrieval over a meeting transcript. The algorithm scores 59% accuracy for passage retrieval, while random guessing is below 1%, but only scores 60% on combined retrieval and question discrimination, for which humans reach 70%–80% and the baseline is 50%. The algorithm clearly outperforms humans for speed, at less than 1 second per question, vs. 1.5–2 minutes per question for humans. The degradation on ASR compared to manual transcripts still yields lower but acceptable scores, especially for passage identification. Automatic QA thus appears to be a promising enhancement to meeting browsers used by humans, as an assistant for relevant passage identification.

Index Terms: question answering, passage retrieval, meeting browsers, evaluation, meeting recording

1. Introduction

Accessing the content of multimedia recordings remains a challenge for current systems despite significant progress in component technologies, such as automatic speech recognition (ASR), speaker diarization, or summarization. Technology progress does not guarantee that meeting browsers using such components will actually be more efficient in helping users to access specific information in meeting recordings.

This paper focuses on tools that facilitate access to specific bits of information from a meeting, as opposed to abstracting information over an entire meeting. The paper describes an automatic question answering (QA) system designed to pass the Browser Evaluation Test (BET) [1, 2], an evaluation method originally intended for human users of a browser. The scores of humans using meeting browsers on the BET task, in terms of speed and precision, are compared with those of our automatic QA system. The results demonstrate the utility of automatic QA techniques as an assistant tool for meeting browsers.

The paper is organized as follows. Section 2 briefly describes the BET task and evaluation protocol. Section 3 describes the automatic QA system, which operates in two stages: passage identification, followed by true/false statement discrimination. Section 4 gives the scores of our QA system in several conditions, and compares them with those of humans using meeting browsers. This comparison shows that humans still 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. The Browser Evaluation Test (BET)

The main quality aspects to be evaluated for interactive software are often categorized as *effectiveness* – the extent to which a browser helps users to accomplish their task, *efficiency* – the speed with which the task is accomplished, and *user satisfaction*. Systematic evaluation of QA systems started with the TREC-8 QA task in 1999 [3], using a database of 1,337 questions from multiple sources, 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 [4]. The 2003 QA track had a passage retrieval task, which inspired also our approach. An evaluation task for interactive QA was also proposed at iCLEF, the Interactive track of the Cross-Language Evaluation Forum [5]; systems-plus-humans were evaluated for accuracy over a large set of questions defined by the experimenters.

The Browser Evaluation Test (BET) [1, 2] is a procedure to collect browser-independent 'questions' about a meeting and to use them for evaluating a browser's capacity to help humans answering them. The questions are in fact pairs of parallel true/false statements which are constructed by neutral 'observers' that view a meeting, write down 'observations of interests' about the meeting, and then create the false counterpart of each observation. Experimenters then gather similar observations into groups. The importance of the group is derived from the observers' own rating of importance and from the size of each group. Examples of the first most important observations (true/false statements) for IB4010, respectively IS1008c, are:

- "The group decided to show The Big Lebowski" *vs.* "The group decided to show Saving Private Ryan".
- "According to the manufacturers, the casing has to be made out of wood" *vs.* "According to the manufacturers, the casing has to be made out of rubber".

Three meetings from the AMI Corpus [6], in English, were selected for the observation collection procedure: IB4010, IS1008c, and ISSCO-Meeting_024, 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 containing the answers is considerably smaller here.

To answer the BET questions, human subjects are required to view the pairs of BET statements in sequence and to decide, using a meeting browser, which statement from the pair is true, and which one is false. The time allowed for each meeting is typically half the duration of the meeting: that is, ca. 24 min. are allowed for IB4010, and 13 min. for IS1008c. The performance of subjects-plus-browsers is measured by the precision of the

*Work performed while at Idiap Research Institute.

answers, which is related to *effectiveness*, and by their speed, which is related to *efficiency*.

The BET results of half a dozen meeting browsers are discussed in a technical report [7]. Average time per question varies from about 1.5 minutes to 4 minutes (with no prior training); most subjects-plus-browsers require on average ca. 2 minutes per question. Precision – against a 50% baseline in the BET application – generally stays in the 70%–80% range, with highest values reaching 90%. The standard deviations are somewhat smaller for precision than for speed, but for both metrics individual performance varies substantially.

3. Automatic BET Question Answering

We have designed a question answering system aimed at discriminating between pairs of BET statements using manual or automatic (ASR) meeting transcripts. The system has an architecture that is inspired by current QA systems [8], with a number of specificities due to the nature of the data and of the task.

The system proceeds in three stages. The first stage is the pre-processing of the pair of BET questions (true/false parallel statements) and of the meeting transcript. The second stage aims to identify separately for each of the questions the passage of the transcript that is most likely to contain the answer to it, using a complex score based on lexical similarity. The third stage compares the two BET statements based on the paragraph found for each question, and hypothesizes which one is true and which one is false.

3.1. Lexical Pre-processing

The transcript and the questions are first prepared for the lexical matching procedure used in passage identification. Initially, abbreviated words are converted into full forms (*we've* → *we have*), numeric forms into text forms (*34* → *thirty four*), upper case letters into lower case ones, and punctuations and stopwords are removed. Apart from frequent function words such as prepositions, adverbs or conjunctions, the list of stopwords includes many pronouns (personal, possessive, and demonstrative) because BET statements are generally formulated using indirect speech with third person pronouns, while the transcript contains utterances in direct form, with first and second person pronouns, leading to lexical mismatches if these are not removed.

The remaining words from the BET question are lemmatized using WordNet, and stemmed using the Snowball implementation of Porter's stemmer. The words from the meeting transcript are processed in the same way, and the name of the speaker of each utterance is included in the data, as names are often quoted in the BET questions. In addition, to each word of the transcript is associated a list of its synonyms from WordNet that have the same part of speech (this is determined using the QTag POS tagger). The reason synonyms are added only to the transcript, and not to the BET questions, is that adding them for both question and transcript words would significantly decrease the predictive power of the lexical matching.

3.2. Passage Identification

Unlike many cases in QA evaluation in which the size of the passage to be retrieved is fixed, here the goal is to retrieve a passage from the transcript that most likely will help to discriminate between the two BET statements, regardless of its size – a very large passage is of course likely not to be helpful. For that, a window of fixed size is moved over the entire transcript, in steps of fixed size, and for each position a lexical similarity score between the words in the window and the words of the candidate BET question is computed. The window that yields the highest score is the returned passage. The score is computed from the number of words from the passage that match words in the candidate question, as in [8]. However, this method is particularized here for the BET task on meetings as follows:

1. If a word from the question matches the name of the speaker of the passage, then it receives the highest possible score value (e.g., a value of 4.0).
2. If a word from the question matches a word from the passage (in terms of lemmas), and this word is spoken by a speaker mentioned in the question, then it receives the second highest score value (e.g., 2.5).
3. Otherwise, if a word from the question matches a word from the passage (lemmas), then it receives the "normal" score value (e.g., 1.0).
4. If a word (lemma) from the question matches one of the synonyms of a word from the passage, then it receives a low score value (e.g., 0.5).

The numeric values listed above were set by the authors based on intuition about the importance of each matching, and might not be optimal for this task. No automatic optimization (statistical learning) was attempted, because the amount of data was insufficient for training and test.

The total score of the passage is the sum of the scores for each matching word. If a word appears several times in the question and/or in the passage, then it is only allowed to count a number of matchings equal to the lowest of the two numbers of occurrences of the word, by discarding pairs of matching words as they are matched.

To find the best matching paragraph, the system must sometimes choose between paragraphs with the same score. In this case, it applies a second time the method described above, but using bigrams of words instead of individual words. If two paragraphs still have the same score, trigrams of words are used, and so on until a difference is found (or a random draw if not). A possible heuristic that has yet to be explored is to favor matching paragraphs that are toward the end of a meeting, as this is the place where "important" ideas are more likely to appear.

3.3. Discrimination of True/False Statements

If a retrieved passage is indeed the discriminating one for the true or the false candidate, then inspiration from work on entailment could be used to find the true statement. The method used here is simpler, and is based on the value of the matching score computed above: the true statement is considered to be the one with the highest matching score. If the best passage scores using word matching are equal for the two candidate statements, then the position of the matched words in the text is used to recompute the scores, favoring passages where matched words are in the same order in both transcript and question. If scores are still the same, bigrams of words are used, and so on.

3.4. Optimization of Parameters

Most of the parameters of the algorithm could be adjusted using statistical learning (optimization) if enough data were available. Given the relatively low number of BET questions available, for two transcribed meetings, we only experimented with brute-force optimization of the window size and window step in the QA algorithm above.

Table 1: *Accuracy of passage retrieval and of true/false statement discrimination for the two BET meetings. Standard deviation (stdev) is computed using 5-fold cross-validation.*

	Passage retrieval				Statement discrimination			
	IB4010		IS1008c		IB4010		IS1008c	
Condition	**Accuracy**	**Stdev**	**Accuracy**	**Stdev**	**Accuracy**	**Stdev**	**Accuracy**	**Stdev**
Random	0.0033	n/a	0.0078	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 matching + speakers	0.55	0.14	0.62	0.16	0.57	0.06	0.64	0.18

Quite early in the development of the algorithm, it appeared useful to have a window size and step which are proportional to the length of a question rather than fixed at the same value for all questions. Using 5-fold cross-validation, all sizes from 1 to 13 times the length of a BET question for windows and steps were tested to find optimal values (the evaluation metric is stated in the next section, 4.1). The results point to a set of optimal values rather than a unique value, and the lowest ones for window size are: for IB4010, 10 times the question size, with a step of 3 times the question size. For IS1008c, these values are respectively 12 and 1, but a window size of 4 times the question length is nearly optimal too, and was selected because a narrower window is more helpful for discrimination.

4. Results for Automatic BET QA

4.1. Evaluation Methods

In order to assess the correctness of a retrieved passage, this is compared to a reference passage annotated by hand, which is the minimal passage that is sufficient to discriminate a BET pair of true/false statements. If the retrieved passage and the correct one have a non-empty intersection (at least one word), then the retrieved passage is considered to be correct.

To assess the correctness of discrimination (second processing stage), it is of course sufficient to check whether the true statement was correctly found or not.

There are 116 pairs of true/false statements for IB4010 and 50 pairs for IS1008c, because only those that were actually shown to humans (after cleaning of the list by BET experimenters) were used in our experiments.

To better assess the scores that follow, we categorized and counted BET questions of two types. The "simple" questions are those that have an explicit answer in some passage, while the "deductive" ones are those the require some reasoning, based on one or more passages, in order to find the answer. Our labeling showed that for IB4010, 74 BET questions out of 116 (64%) are simple and the remaining 42 (36%) are deductive. For IS1008c, 46 questions out of 50 (92%) are simple and only 4 (8%) are deductive. These figures provide an upper margin for the performance of our system, which cannot be expected to answer deductive questions at this stage, though it still can find the corresponding passages.

The baseline score for passage retrieval can be defined by random draw, and is thus dependent on the size of the window and its step. With the current values of these sizes, the likelihoods of finding the correct passage by chance are less than 1% for each meeting. The baseline score for true/false discrimination is 50%.

4.2. Results on Manual Transcripts

Results for the two processing stages of the automatic BET QA algorithm are given in Table 1 above, for three variants of the algorithm: (1) allowing only unigram matching when computing the similarity score, with no weighting of speaker-specific words; (2) with N-gram matching and still no weighting; and (3) with the additional weighting of matched words spoken by a speaker mentioned in the question, as explained in Section 3.2 above, second bullet point.

The passage retrieval component obtains very good results compared with the chances of randomly locating the correct passage, with accuracy scores of 0.55 ± 0.14 for IB4010 and 0.62 ± 0.16 for IS1008c[1]. These results also compare favorably with overall human-plus-browser precision scores, shown to be in the 70%–80% range, though they remain clearly below them. As for speed, the automatic QA system is of course much faster than humans-plus-browsers, requiring less than 1 second per question (vs. 2–4 minutes).

When combined with the question discrimination, the performance degrades at 0.57 ± 0.06 accuracy for IB4010 and 0.64 ± 0.18 for IS1008c (with respect to the 50% baseline). A more informative comparison is done by noting that a perfect discrimination module would still be clueless on passages that are wrongly identified (45% and 38%), and so its score could only reach at most 77.5% for IB4010 and 81% for IS1008c[2]. The fact that the actual scores are clearly lower than these theoretical values shows that the algorithm needs improvement for this stage.

In order to answer the question: "do humans and our QA system have difficulties on the same BET statements?", a detailed comparison, by question, of the automatic scores with those of humans using the TQB browser [2] is shown in Table 2. For humans using TQB, the tables show the scores of a group of 14 people without training, and the scores of an equivalent group after training on one meeting [2, 7]. The amount of available data does not allow a full statistical study of the differences between humans and the QA system based on correlations, but the observation of the first eight questions for each meeting seem to suggest some correlation between the what humans and the QA system find difficult, e.g. question #1 for IB4010, and questions #5 and #6 for IS1008c.

[1] Confidence intervals at the 95% level are obtained through 5-fold cross validation.

[2] That is, if question discrimination was perfect, it would work on the correctly retrieved passages, and would reach 50% accuracy on the others, so the expected scores would be $0.55 * 100\% + 0.45 * 50\% = 77.5\%$ for IB4010 and respectively $0.62 * 100\% + 0.38 * 50\% = 81\%$ for IS1008c.

Table 2: *Human (with TQB browser) vs. automatic results for the first eight questions of the meetings (P: accuracy of passage retrieval; D: accuracy of statement discrimination).*

Question number	Precision of humans no training	with training	QA System P	D
IB4010: 1	0.93	0.71	0	0
2	0.93	1.00	1	1
3	0.71	1.00	1	1
4	0.86	0.86	1	1
5	1.00	0.93	0	1
6	0.93	1.00	1	1
7	0.93	0.71	1	1
8	0.71	0.79	1	1
Average	0.88	0.88	0.75	0.88
IS1008c: 1	0.86	0.93	1	1
2	0.67	0.86	1	1
3	0.82	0.93	1	1
4	0.89	0.93	1	1
5	0.63	0.69	1	0
6	0.67	0.73	0	0
7	1.00	0.82	1	0
8	0.67	0.64	0	1
Average	0.77	0.81	0.75	0.63

4.3. Results on ASR and Summaries

The automatic BET QA system was also applied to the diarized ASR from the same meetings which is available with the AMI Corpus [6], as a sample of the output of the AMI ASR system [9]. As expected, the scores decrease on ASR with respect to the manual transcript, but remain in a similar range.

For IB4010, passage retrieval accuracy drops to 0.46 ± 0.13 from 0.55 ± 0.14, and discrimination accuracy drops to 0.52 ± 0.09 from 0.57 ± 0.06 (thus getting very close to the baseline score of 50%). For IS1008c, passage retrieval drops to 0.60 ± 0.33 from 0.62 ± 0.16, and discrimination drops to 0.56 ± 0.19 from 0.64 ± 0.18. Our QA system appears thus to be robust with respect to the quality of the ASR. Especially passage retrieval scores remain at levels that are quite high above the baseline: the system finds the right passage for more than half of the BET questions.

We also assessed the degradation when automatic summaries [10] over the ASR are used instead of the full ASR. The goal is first to test the robustness of our system, but also to assess summarization quality. We indeed believe that this approach could provide an indirect measure of the quality of a summary: the higher the quality of the summary, the lower the degradation when shorter and shorter summaries are used. We compared extractive summaries of various sizes with a random extract and with a gold standard summary from the AMI Corpus, but the initial results did not show significant differences between the various conditions.

5. Conclusion and Future Work

This paper has described an automatic system for passing a QA-based browser evaluation task, the BET, and has compared its results with those of human subjects using meeting browsers. Although significantly above the baseline, the performance of automatic QA appeared to be quite below that of humans for the overall precision on the entire task, i.e. the discrimination of true vs. false statements. However, the meeting-specific QA algorithm introduced here displays two clear advantages over humans: it retrieves the correct passage for more than 50% of the questions, in a very short time, smaller than 1 second per question.

These results suggest that an improvement to meeting browsers – at least for the fact-finding task – is to enhance them with an automatic passage retrieval function, which points the user from the start to a potentially relevant section of a meeting, depending on the question that was formulated. If the section is not relevant, then the user starts using the meeting browser as before. But if the section is indeed relevant, as in more than half of the cases, then the user can easily reason upon it (if needed) to extract the answer, in a more reliable way than the method proposed here. Future work on entailment could also help to improve the automation of this second stage.

6. 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. The first author was supported by the AMI Training Program during his internship at Idiap (Aug. 2008 – Jan. 2009).

7. References

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

[2] 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.

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

[4] ——, "Overview of the TREC 2003 question answering track," in *TREC 2003 (12th Text REtrieval Conference)*, ser. NIST Special Publication 500-255, Gaithersburg, MD, 2003, pp. 54–68.

[5] 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.

[6] 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.

[7] A. Popescu-Belis, "Comparing meeting browsers using a task-based evaluation method," Idiap Research Institute, Research Report Idiap-RR-11-09, June 2009.

[8] M. Light, G. S. Mann, E. Riloff, and E. Breck, "Analyses for elucidating current question answering technology," *Natural Language Engineering*, vol. 7, no. 4, pp. 325–342, 2001.

[9] T. Hain *et al.*, "The development of the AMI system for the transcription of speech in meetings," in *Machine Learning for Multimodal Interaction II*, ser. LNCS 3869, S. Renals and S. Bengio, Eds. Berlin/Heidelberg: Springer-Verlag, 2006, pp. 344–356.

[10] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Interspeech 2005*, Lisbon, 2005, pp. 593–596.