Overview of QAST 2007

Jordi Turmo[1], Pere Comas[1], Christelle Ayache[2], Djamel Mostefa[2] and Sophie Rosset[3] and Lori Lamel[3]
[1]TALP Research Centre (UPC). Barcelona. Spain
{turmo,pcomas}@lsi.upc.edu
[2]ELDA/ELRA. Paris. France
{ayache,mostefa}@elda.org
[3]LIMSI. Paris. France
{rosset,lamel}@limsi.fr

Abstract

This paper describes QAST, a pilot track of CLEF 2007 aimed at evaluating the task of Question Answering in Speech Transcripts. The paper summarizes the evaluation framework, the systems that participated and the results achieved. These results have shown that question answering technology can be useful to deal with spontaneous speech transcripts, so for manually transcribed speech as for automatically recognized speech. The loss in accuracy from dealing with manual transcripts to dealing with automatic ones implies that there is room for future reseach in this area.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Experimentation, Performance, Measurement

Keywords

Question Answering, Spontaneous Speech Transcripts

1 Introduction

The task of Question Answering (QA) consists of providing short, relevant answers to natural language questions. Most Question Answering research has focused on extracting information from text sources, providing the shortest relevant text in response to a question [4, 5]. For example, the correct answer to the question *How many groups participate in the CHIL project?* is *16*. Whereas the response to the question of *who are the partners in CHIL?* is a list of the partners. This simple example illustrates the two main advantages of QA has over current search engines: first, the input is a natural language question rather a keyword query, and second, the answer provides the desired information content and not a potentially large set of documents or URLs that the user must plow through.

Most of current QA systems handle independent questions and produce one answer to each question, extracted from textual data, for both open domain and limited domain tasks. However, a large portion of human interactions involve spontaneous speech, e.g. meetings, seminars, lectures, telephone conversations, and are beyond the capacities of current text-based factual QA

systems. Most of the recent QA research has been undertaken by natural language groups who have typically applied techniques to written texts, and assume that these texts have a correct syntactic and semantic structure. The grammatical structure of spoken language is different from that of written language, and some of the anchor points used in text processing such as punctuation must be inferred and are therefore error prone. Other spoken language phenomena include disfluencies, repetitions, restarts and corrections. In the case that automatic processing is used to create the speech transcripts, an additional challenge is dealing with the recognition errors. The lecture and interactive meeting data are particularly difficult due to run-on sentences (where the distance between the first part of an utterance and its end one can be very long) and interruptions. Therefore current techniques for text-based QA need substantial adaptation in order to access the information contained in audio data.

This paper provides an overview of a pilot evaluation track at CLEF 2007 for Question Answering in Speech Transcriptions, named QAST. Section 2 describes the principles of this evaluation track. Sections 3 and 4 present the evaluation framework and the systems that participated, respectively. Section 5 shows the results achieved and the main implications. Finally, Section 6 concludes.

2 The QAST task

The objective of this pilot track is to provide a framework in which QA systems can be evaluated when the answers have to be found in spontaneous speech transcripts (manual and automatic transcripts). There are three main objectives to this evaluation:

- Comparing the performances of the systems dealing with both types of transcripts.
- Measuring the loss of each system due to the inaccuracies in state of the art ASR technology.
- Motivating and driving the design of novel and robust factual QA architectures for automatic speech transcripts.

In this evaluation, the QA systems have to return answers found in the audio transcripts to questions presented in a written natural language form. The answer is the minimal sequence of words that includes the correct exact answer in the audio stream. For the purposes of this evaluation, instead of pointers in the audio signal, the recognized words covering the location of the exact answer have to be returned. For example, consider the question *which organisation has worked with the University of Karlsruhe on the meeting transcription system?*, and the following extract of an automatically recognized document:

> *breath fw and this is , joint work between University of Karlsruhe and coming around so fw all sessions , once you find fw like only stringent custom film canals communicates on on fw tongue initials .*

corresponding to the following exact manual transcript:

> *uhm this is joint work between the University of Karlsruhe and Carnegie Mellon, so also here in these files you find uh my colleagues and uh Tanja Schultz.*

The answer found in the manual transcript is *Carnegie Mellon* whereas in the automatic transcript it is *coming around*. This example illustrates the two principles that guide this track:

- The questions are generated considering the exact information in the audio stream regardless of how this information is transcribed, because the transcription process is transparent to the user.

- The answer to be extracted is the minimal sequence of words that includes the correct exact answer in the audio stream (i.e., in the manual transcripts). In the above example, the answer to be extracted from the automatic transcript is *coming around*, because this text gives the start/end pointers to the correct answer in the audio stream.

Four tasks have been defined for QAST:

- T1: QA in manual transcriptions of lectures.
- T2: QA in automatic transcriptions of lectures.
- T3: QA in manual transcripts of meetings.
- T4: QA in automatic transcriptions of meetings.

3 Evaluation protocol

3.1 Data collections

The data for the QAST pilot track consists of two different resources, one for dealing with the lecture scenario and the other for dealing with the meeting scenario:

- The CHIL corpus[1]: it consists of around 25 hours (around 1 hour per lecture) both manually and automatically transcribed (LIMSI produced the ASR transcriptions with around 20% of word error rate -WER- [2], while the manual ones were done by ELDA). In addition, the set of lattices and confidences for each lecture has been provided. The domain of the lectures is *speech and language processing*. The language is European English (mostly spoken by non native speakers). Lectures have been provided with simple tags. Seminars are formatted as plain text files (ISO-8859-1) [3].
- The AMI corpus[2]: it consists of around 100 hours (168 meetings) both manually and automatically transcribed (the Univeristy of Edimburgh produced the ASR trasncripts with around 38% of WER [1]). The domain of this meetings is *design of television remote control*. The language is European English. Meetings (as lectures) have been produced with simple tags. Meetings are formatted as plain text files (ISO-8859-1).

3.1.1 Questions and answer types

For each one of the scenarios, two sets of questions will be provided to the participants:

- Development set (1 February 2007) :
  - Lectures: 10 seminars and 50 questions.
  - Meetings: 50 meetings and 50 questions.
- the Evaluation set (18 June 2007):
  - Lectures: 15 seminars and 100 questions.
  - Meetings: 118 meetings and 100 questions.

Question sets have been formatted as plain text files, with one question per line as defined in the Guidelines[3]. All the questions in the QAST task are factual questions, whose expected answer is a Named Entity (person, location, organization, language, system, method, measure, time, color, shape and material). No definition questions have been proposed. The two data collections (CHIL

[1] http://chil.server.de
[2] http://www.amiproject.org
[3] http://www.lsi.upc.edu/~qast

and AMI corpus) were first tagged with Named Entities. Then, an English native speaker created questions for each NE tagged session. So each answer is a tagged Named Entity.

An answer is basically structured as an [answer-string, document-id] pair, where the answer-string contains nothing more than a complete and exact answer (a Named Entity) and the document-id is the unique identifier of a document that supports the answer. There are no particular restrictions on the length of an answer-string (which is usually very short), but unnecessary pieces of information will be penalised, since the answer will be marked as non-exact. Assessors will focus mainly on the responsiveness and usefulness of the answers.

3.2 Human judgement

The files submitted by participants have been manually judged by native speaking assessors. Assessors considered correctness and exactness of the returned answers. They have also checked that the document labelled with the returned docid supports the given answer. One assessor evaluated the results. Then, another assessor manually checked each judgement evaluated by the first one. Any doubts about an answer was solved through various discussions.

To evaluate the data, assessors used an evaluation tool developed in Perl (at ELDA) named QASTLE[4]. A simple interface permits easy access of the question, the answer and the document associated with the answer (all in one window only).

For T2 and T4 (QA on automatic transcripts) the manual transcriptions were aligned to the automatic ASR outputs to find the answer in the automatic transcripts. The alignments between the automatic and the manual transcription were done using time information for most of the seminars and meetings. Unfortunately for some AMI meetings time information were not available and only word alignments were used.

After each judgement the submission files have been modified. A new element appears in the first column: the answer's evaluation (or judgement). The four possible judgements (also used at TREC[5]) correspond to a number ranging between 0 and 3:

- 0 correct: the answer-string consists of the relevant information (exact answer), and the answer is supported by the returned document.
- 1 incorrect: the answer-string does not contain a correct answer or the answer is not responsive.
- 2 non-exact: the answer-string contains a correct answer and the docid supports it, but the string has bits of the answer missing or is longer than the required length of the answer.
- 3 unsupported: the answer-string contains a correct answer but the docid does not support it.

3.3 Measures

The two following metrics used in CLEF have been used in the QAST evaluation:

1. Mean Reciprocal Rank (MRR) measures how well ranked is the right answer, as defined in Section 2, in the list of 5 possible answers in average.
2. Accuracy: The fraction of correct answers ranked in the first position in the list of 5 possible answers.

[4] http://www.elda.org/qastle/

4 Submitted runs

A total of five groups from five different countries submitted results for one or more of the proposed QAST tasks. Due to various reasons (technical, financial, etc.), three other registered groups were not be able to submit any results.

The five participating groups are the following:

- CLT, Center for Language Technology, Australia;
- DFKI, Germany;
- LIMSI, Laboratoire d'Informatique et de Mécanique des Sciences de l'Ingénieur, France;
- TOKYO, Tokyo Institute of Technology, Japan;
- UPC, Universitat Politècnica de Catalunya, Spain.

Five groups participated in both T1 and T2 tasks (CHIL corpus) and three groups participated in both T3 and T4 tasks (AMI corpus).

The participants could submit up to 2 submissions per task and up to 5 answers per question. The systems used in the submissions are described in Table 1. In total, 28 submissions were evaluated: 8 submissions from 5 participating sites for T1, 9 submission files from 5 different sites for T2, 5 submissions from 3 participants for T3 and 6 submissions from 3 participants for T4. The lattices provided for task T2 were not finally used by any participant.

<table>
<tr><th>System</th><th>Enrichment</th><th>Question classification</th><th>Doc/Pass Retrieval</th><th>Answer Extraction</th><th>NERC</th></tr>
<tr><td>clt1</td><td rowspan="2">words and NEs</td><td rowspan="2">hand-crafted patterns</td><td rowspan="2">pass. ranking based on word similarities between pass. and query</td><td rowspan="2">candidate ranking based on frequency and the NER confidence</td><td>hand-crafted patterns, gazeetters and ME models</td></tr>
<tr><td>clt2</td><td>No ME models</td></tr>
<tr><td>dfki1</td><td>words and NEs</td><td>hand-crafted sint.-sem. rules</td><td>Lucene</td><td>candidate ranking based on frequency</td><td>gazeeteers and not tuned statistical models</td></tr>
<tr><td>limsi1</td><td rowspan="2">words and NEs</td><td rowspan="2">hand-crafted patterns</td><td>pass. ranking based on hand-crafter back-off queries</td><td rowspan="2">candidate ranking based on frequency, keyword distance and retrieval confidence</td><td rowspan="2">hand-crafted patterns</td></tr>
<tr><td>limsi2</td><td>cascaded doc/pass ranking based on search descriptors</td></tr>
<tr><td>tokyo1</td><td rowspan="2">words</td><td rowspan="2">non-linguistic statistical multi-word model</td><td>pass. retrieval with interpolated doc/pass statistical models</td><td rowspan="2">candidate ranking based on statistical multi-word model</td><td rowspan="2">no</td></tr>
<tr><td>tokyo2</td><td>addition of word classes to the statistical models</td></tr>
<tr><td>upc1</td><td>words, NEs lemmas and POS</td><td rowspan="2">perceptrons</td><td>pass. ranking based on iterative query relaxation</td><td>candidate ranking based on keyword distance and density</td><td rowspan="2">hand-crafted patterns, gazeetters and perceptrons</td></tr>
<tr><td>upc2</td><td>also phonetics</td><td colspan="2">addition of approximated phonetic matching</td></tr>
</table>

Table 1: Systems that participated in QAST

5 Results

The results for the four QAST tasks are presented in tables 2, 3, 4 and 5. Due to some problems (typo, answer type) some questions have been deleted from the scoring results in tasks T1, T2 and T3. In total, the results have been calculated on the basis of 98 questions for tasks T1 and

T2, and 96 for T3. In addition, and due to also missing time information at word level for some AMI meetings, seven questions have been deleted from the scoring results of T4. The results for this task have been calculated on the basis of 93 questions.

System	# Questions	#Correct answers	MRR	Accuracy
clt1_t1	98	16	0.09	0.06
clt2_t1	98	16	0.09	0.05
dfki1_t1	98	19	0.17	0.15
limsi1_t1	98	43	0.37	0.32
limsi2_t1	98	56	0.46	0.39
tokyo1_t1	98	32	0.19	0.14
tokyo2_t1	98	34	0.20	0.14
upc1_t1	98	54	0.53	0.51

Table 2: Results for T1 (QA on CHIL manual transcriptions)

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t2	98	13	0.06	0.03
clt2_t2	98	12	0.05	0.02
dfki1_t2	98	9	0.09	0.09
limsi1_t2	98	28	0.23	0.20
limsi2_t2	98	28	0.24	0.21
tokyo1_t2	98	17	0.12	0.08
tokyo2_t2	98	18	0.12	0.08
upc1_t2	96	37	0.37	0.36
upc2_t2	97	29	0.25	0.24

Table 3: Results for T2 (QA on CHIL automatic transcriptions)

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t3	96	31	0.23	0.16
clt2_t3	96	29	0.25	0.20
limsi1_t3	96	31	0.28	0.25
limsi2_t3	96	40	0.31	0.25
upc1_t3*	95	23(27)	0.22(0.26)	0.20(0.25)

Table 4: Results for T3 (QA on AMI manual transcriptions). *Due to a bug with the output format script, UPC asked to the assessors to reevaluate their unique run for T3. The results in brackets must be regarded as a non official run.

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t4	93	17	0.10	0.06
clt2_t4	93	19	0.13	0.08
limsi1_t4	93	21	0.19	0.18
limsi2_t4	93	21	0.19	0.17
upc1_t4	91	22	0.22	0.21
upc2_t4	92	17	0.15	0.13

Table 5: Results for T4 (QA on AMI manual transcriptions)

The results are very encouraging. First, the best result in accuracy achieved in tasks involving manual transcripts (0.51 for task T1) is closed to the best two results for factual questions in TREC 2006 (0.58 and 0.54), in which monolingual English QA was evaluated. Second, this behaviour is also observed in average: the accuracy in average achieved in tasks T1 and T3 is 0.22, which is

comparable with 0.18 achieved in TREC 2006. Although no direct comparisons between QAST and TREC are possible due to the use of different data, questions and answer types, these facts show that QA technology can be useful to deal with spontaneous speech transcripts.

Finally, the accuracy values are 0.22 and 0.15 in average for the tasks involving lectures (T1 and T2, respectively), and 0.21 and 0.14 for those involving meetings (T3 and T4, respectively). These values show that the accuracy decreases in average more than 36% when dealing with automatic transcripts. The reduction of this difference between accuracy values have to be taken as a main goal in the future research.

6 Conclusion

In this paper, we have described the QAST 2007 (Question Answering in Speech Transcripts) task. A set of five groups participated in this track with a total of 28 submitted runs among four specific tasks. In general, the results achieved show that, first, QA technology can be useful to deal with spontaneous speech transcripts, and second, the loss in accuracy when dealing with automatically transcribed speech is high. These results are very encouraging and suggest that there is room for future research in this area.

Future work aims at including in the evaluation framework other languages than English, oral questions, and other question types different than factual ones.

Acknowledgments

We are very grateful to Thomas Hain from the University of Edimburgh, who provide us with the AMI transcripts automatically generated by their ASR. This work has been jointly funded by the European Commission (CHIL project IP-506909), the Spanish Ministry of Science (TEXTMESS project) and the LIMSI AI/ASP RITEL grant.

References

[1] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The ami system for the transcription of meetings. In *Proceedings of ICASSP'07*, 2007.

[2] L. Lamel, G. Adda, E. Bilinski, and J.-L. Gauvain. Transcribing lectures and seminars. In *Proceedings of Interspeech'05*, 2005.

[3] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnvmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, L Bernardin, and C. Rochet. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *to appear in Language Resources and Evaluation Journal*, 2007.

[4] C. Peters, P. Clough, F.C. Gey, J. Karlgren, B. M;agnini, D.W. Oard, M. de Rijke, and M. Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval*. Springer-Verlag., 2006.

[5] E.M. Voorhees and L.L. Buckland, editors. *The Fifteenth Text Retrieval Conference Proceedings (TREC 2006)*, 2006.