Evaluation Protocol and Tools for Question-Answering on Speech Transcripts

N. Moreau[1], O. Hamon[1], D. Mostefa[1], S. Rosset[2], O. Galibert[2(*)], L. Lamel[2], J. Turmo[3], P. R. Comas[3], P. Rosso[4], D. Buscaldi[4], K. Choukri[1]

[1]ELDA (Evaluations and Language resources Distribution Agency), [2]LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), [(*)] now at LNE (Laboratoire National de métrologie et d'Essais), [3]UPC (Universitat Politècnica de Catalunya), [4]UPV (Universitat Politècnica de València)

Address: ELDA, 55-57 rue Brillat Savarin, 75013 Paris, France.

E-mails: {moreau;hamon;mostefa}@elda.org, sophie.rosset@limsi.fr, olivier.galibert@lne.fr, lamel@limsi.fr, {turmo;pcomas]@lsi.upc.edu,{prosso;dbuscaldi}@dsic.upv.es, choukri@elda.org

Abstract

Question Answering (QA) technology aims at providing relevant answers to natural language questions. Most Question Answering research has focused on mining document collections containing written texts to answer written questions. In addition to written sources, a large (and growing) amount of potentially interesting information appears in spoken documents, such as broadcast news, speeches, seminars, meetings or telephone conversations. The QAST track (Question-Answering on Speech Transcripts) was introduced in CLEF to investigate the problem of question answering in such audio documents. This paper describes in detail the evaluation protocol and tools that were developed for the CLEF-QAST evaluation campaigns that have taken place between 2007 and 2009.

1. Introduction

Question Answering (QA) technology aims at providing relevant answers to natural language questions. Most Question Answering research has focused on mining document collections containing written texts to answer written questions [1]. In addition to written sources, a large (and growing) amount of potentially interesting information appears in spoken documents, such as broadcast news, speeches, seminars, meetings or telephone conversations. The QAST track (Question-Answering on Speech Transcripts) was introduced in CLEF to investigate the problem of question answering in such audio documents.

This paper describes in detail the evaluation protocol and tools that were developed for the CLEF-QAST evaluation campaigns that have taken place in 2007 [2], 2008 [3] and 2009 [4]. The QAST Evaluation Package 2007-2009 resulting from these evaluation campaigns is also introduced.

2. Evaluation Data and Tasks

2.1. Data Collections

Along the years, participants to QAST campaigns were proposed different evaluation scenarios and tasks, each one involving a different data set:

(1) CHIL corpus: lectures in English on topics related to *"speech and language processing"*,

(2) AMI corpus: meetings in English about *"design of television remote control"*,

(3) ESTER corpus: French broadcast news,

(4) EPPS-EN corpus: European Parliament debates in English,

(5) EPPS-ES corpus: European Parliament debates in Spanish.

For each corpus two types of transcriptions were available and had to be processed:

- *Manual Transcriptions*: the exact manual transcriptions (including speech disfluencies) of the original audio documents were done at ELDA[1].
- *Automatic (or ASR) Transcriptions*: automatic transcriptions of the data sets were also available. They were produced by multiple automatic speech recognition (ASR) engines that have been developed in the context of European and national projects: the CHIL project [5][6] for corpus (1), the AMI project [7] for corpus (2), the ESTER project [8] for corpus (3), and the TC-STAR project [9] for (4) and (5).

Table 1 gives more details on the evaluation corpora used in QAST from 2007 to 2009.

Corpus	Lang.	Description	Transcripts	WER	Campaigns
CHIL	EN	25 lectures (~25h)	manual	-	2007, 2008
			1 set ASR	20%	2007, 2008
AMI	EN	168 meetings (~100h)	manual	-	2007, 2008
			1 set ASR	38%	2007, 2008
ESTER	FR	18 BN shows (~10h)	manual	-	2008, 2009
			3 sets ASR	A: 11.9% B: 23.9% C: 35.4%	2008, 2009
EPPS	EN	6 sessions (~3h)	manual	-	2008, 2009
			3 sets ASR	A: 10.6% B: 14.0% C: 24.1%	2008, 2009
EPPS	ES	6 sessions (~3h)	manual	-	2008, 2009
			3 sets ASR	A: 11.5% B: 12.7% C: 13.7%	2008, 2009

Table 1. QAST evaluation data sets, with word error rates (WER) of automatic transcription corpora (ASR).

[1]ELDA: http://www.elda.org

2.2. Question Sets

Each year, 2 new sets of questions were created from each evaluation corpus:

- Development set: 50 questions,
- Test set: 100 questions.

In 2007, only factual questions were created, based on 10 types of named entities: person, location, organization, language, system, measure, time, color, shape, material.

In 2008, definition questions were introduced (around 75% factual and 25% definition questions in each set) based on 4 types: person, organization, object, other.

In 2009, the same types of questions were used, but a new question collection protocol was designed. In the previous years, written questions were created by hand from each corpus by a single reader. In 2009, spontaneous oral questions were recorded by several speakers just after they had read pieces of texts extracted from the corpora [10]. Oral questions were transcribed (including speech disfluencies). A clean written version of these transcripts was produced afterwards, resulting in two types of questions for each set:

- Spontaneous oral questions (i.e. their transcripts)
- Plus their "canonical" written equivalents.

Example of transcription of a 2009 spontaneous oral question:

When did the bombing of Fallujah t() take took place?

and its written equivalent:

When did the bombing of Fallujah take place?

2.3. Evaluation Tasks

Based on these data and question sets, different evaluation tasks were proposed each year to the participants:

- QA on the manual transcription of each evaluation corpus,
- QA on the different sets of ASR transcriptions assigned to each evaluation corpus.

In 2009, participants could use spontaneous oral questions (in addition to written questions) to test their QA system both on manual and ASR transcriptions.

3. Submission

3.1 Submission Procedure

Each year, QAST participants were first sent the training dataset (texts and questions) prior to the start of the evaluation, in order to train their systems with the required question types. Then, as soon as the evaluation campaign was started, they received test collections and question sets. They had one week to return their QA systems' answers to the evaluation agency (ELDA).

3.2 Submission Format

The required answer format was basically structured as an [*answer-string, document-id*] pair, where the answer-string contains nothing more than a complete and exact answer and the document-id is the unique identifier of a document that supports the answer.

There were no particular restrictions on the length of an answer string, but participants were aware that unnecessary pieces of information would be penalized with the answer being assessed as "*inexact*".

A run had to be submitted as a single text file containing one line per answer, with the following format:

```
<question-id>  <run-id>  <document-id>
<answer-string>  <ranking>  <score>  <starttime>
<endtime>
```

where:

- <*question-id*> is the question identification number,
- <*run-id*> identifies the submitted run (participant, sub-task),
- <*document-id*> contains the name of the document where the answer was found (or a blank if no answer was found),
- <*answer-string*> contains the answer-string (or '*NIL*' if no answer was found),
- <*ranking*> is the answer's rank (it was possible to submit up to 5 answers to a same question),
- <*score*> (or confidence score) is a mandatory score-value per answer,
- <*starttime*> and <*endtime*> are mandatory only if automatic transcripts are used and give the position of the answer in the signal (extracted from the ASR transcription files).

Examples:

Questions:

```
...
38 Which university is located in Dallas?
39 What language has the most important economic
impact?
...
```

Answers in manual transcriptions

```
...
38 limsi1_t1a ISL_20050112 southern methodist
university 1 0.76
38 limsi1_t1a NIL 2 0.68
39 limsi1_t1a ISL_20050420 english 1 0.52
39 limsi1_t1a ISL_20050420 english 2 0.50
39 limsi1_t1a ISL_20050112 dutch 3 0.42
...
```

Answers in automatic transcriptions

```
...
38 limsi1_t1b ISL_20050420 Southern at the
University 1 0.76 94.340 95.310
38 limsi1_t1b NIL 2 0.68
39 limsi1_t1b ISL_20050420 English 1 0.52 551.800
552.120
39 limsi1_t1b ISL_20050420 English 2 0.50 1263.920
```

```
1264.320
39 limsi1_t1b ISL_20050112 Dutch 3 0.42 836.400
837.020
...
```

4. Assessment of Answers

4.1 Assessment for Manual Transcriptions

The submitted files were manually judged by two native-speaking assessors. Assessors had to consider *correctness* (i.e. responsiveness) and *exactness* (i.e. the quantity of information) of the returned answers. They also checked that the document labelled with the returned *document-id* supports the given response. Each [*answer-string*, *document-id*] pair was assessed and marked with one of the following judgments (that have also been used at TREC):

- *Right* (R): the answer-string consists of the relevant information (exact answer), and the answer is supported by the returned document,
- *Wrong* (W): the answer-string does not contain a correct answer or the answer is not responsive,
- *Unsupported* (U): the answer-string contains a correct answer but its *document-id* does not support it,
- *Inexact* (X): the answer-string contains a correct answer and the *document-id* supports it, but the string has bits of the answer missing or is longer than the required length of the answer.

Assessors used a graphical interface developed at ELDA and named QASTLE [2]. The QASTLE tool aims at facilitating the evaluation of question-answering systems for human judges.

QASTLE has already been used successfully for past evaluation campaigns, such as EQueR [11] in France and was specifically redesigned to match the requirements of the QAST CLEF track.

Figure 1 shows a snapshot of QASTLE. The interface can be separated into two mains parts:

- The left side concerns the evaluation of the answers to a question.
- The right side displays the documents corresponding to the answers (i.e. where they have been found).

On top of the left part, the currently selected question is displayed. Below appears the list of all submitted answers, first appearing by default in grey (*not assessed*). When the judge selects an answer by clicking on it, the associated document is displayed on the right part. The answer is then assessed using 4 buttons (*Right, Wrong, Unsupported, Inexact*). Once assessed, the answer's color changes to the color of the chosen assessment button. This allows to quickly visualize the results of all assessments (useful in case of cross verification by a second judge).

The document where the selected answer has been found is displayed in the right window and can be explored thanks to a simple keyword search engine, thus helping towards faster assessment of the answer. The searched keywords are highlighted in the document.

QASTLE displays the pool of all submitted answers (i.e. yielded by all participating QA systems) to the same question. Judges assess them all in a sequential manner and then click on "next question". This pool-based procedure greatly speeds up and enhances the assessment work.

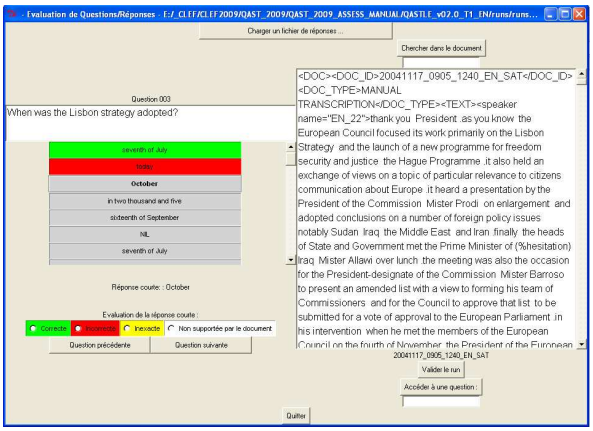

Figure 1. The QASTLE evaluation interface.

Each time a judgment is made, QASTLE automatically inserts the corresponding identification letter (R, W, X, U) at the beginning of the answer line in the submitted file, as follows:

```
(...)
R  099  limsi1_t1a  ISL_20050420 Cambridge 1
   0.92
R  100  limsi1_t1a  ISL_20050420 VTLN       1
   0.89
W  101  limsi1_t1a  NIL  1  0.69
(...)
```

Finally, the assessed files are processed with a script to compute the two following evaluation metrics:

- *Accuracy*: fraction of correct answers ranked in the first position within the list of possible answers.
- *Mean Reciprocal Ranked (MRR)*: reciprocal of the rank of the first correct answer, averaged over all questions.

4.2 Assessment for Automatic Transcriptions

The submitted run using automatic transcriptions had to provide the time slot of the answer, i.e. the timestamps of the beginning and the end of the corresponding word sequence in the transcription.

Based on this time-stamps information, submissions made using ASR transcriptions were evaluated according to a different protocol.

First, prior to the evaluation campaign, reference time slots were created by hand for each set of questions:

- All possible answers of all test questions were manually searched for in the documents of the test collection.

[2] QASTLE: http://elda.org/qastle/

- Each found answer was manually labelled with its actual timestamps in the audio signal (start time and end time).

Then, after the participants had submitted their runs, the assessment procedure consisted in two steps:

- Submitted files were assessed with an automatic script which compared the time slots of submitted answers to the time slots of reference.
- The automatic assessments were finally checked by hand by a human assessor using the QASTLE interface.

In the first pass, a submitted answer is assessed via a script by comparing its hypothesis time slot $[TH_{start};TH_{end}]$ to the time slots of reference $[TR_{start};TR_{end}]$ (there can be several reference slots, since correct answers may appear in different parts of the corpus). The decision procedure implemented in the assessment script is the following:

- If a sufficient overlap is observed between a submitted answer and one of the answers of reference, this answer is tagged as correct. In other words, if there is at least one reference answer $[TR_{start};TR_{end}]$ for which:

$$TR_{start} - \Delta T \leq TH_{start} \leq TR_{start} + \Delta T$$
$$\text{AND} \quad TR_{end} - \Delta T \leq TH_{end} \leq TR_{end} + \Delta T$$

  then the answer contained in $[TH_{start};TH_{end}]$ is set to *R* (correct).

- Else, if there is at least one reference answer $[TR_{start};TR_{end}]$ that overlaps the hypothesis time slot $[TH_{start};TH_{end}]$, then the answer contained in $[TH_{start};TH_{end}]$ is set to *X* (inexact).
- Else, the answer contained in $[TH_{start};TH_{end}]$ is set to *W* (wrong).

The overlap threshold (defined by the delta value ΔT) is derived from word-length statistics. A specific delta value ΔT has been computed beforehand for each of the transcription sets by taking the 95th percentile value in each case. These values are given in Table 2:

ASR Transcripts	WER	ΔT
CHIL	20.0%	610 ms
AMI	38.0%	630 ms
ESTER – A	11.9%	600 ms
ESTER – B	23.9%	630 ms
ESTER – C	35.4%	640 ms
EPPS-EN – A	10.6%	700 ms
EPPS-EN – B	14.0%	680 ms
EPPS-EN – C	24.1%	750 ms
EPPS-ES – A	11.5%	720 ms
EPPS-ES – B	12.7%	700 ms
EPPS-ES – C	13.7%	760 ms

Table 2. Delta values of the transcription sets.

In the second pass (manual checking of automatic assessments), a human assessor had to ensure that each 'correct' (R) or 'inexact' (X) answer could be found in the associated document: if not, it was retagged as 'unsupported' (U). When an answer tagged as 'wrong' (W) or 'inexact' (X) was re-assessed as 'correct' by the assessor, the corresponding time slot was manually adjusted or added in the reference and all runs were reassessed according to the new updated list of reference answers.

4.3 Results

Table 3 gives a very short overview of the results obtained in the three past QAST campaigns (the best accuracy score is given in each case). For QAST 2009 two columns of results are given: the right ones result from using 'oral' questions (i.e. exact transcriptions of spontaneous oral questions), the left one result from using 'written' questions (i.e. their canonical form).

Corpus	Transcr.	Acc. 2007	Acc. 2008	Acc 2009 Written	Acc. 2009 Oral
CHIL	Manual	0.51	0.41	-	-
	ASR	0.36	0.31	-	-
AMI	Manual	0.25	0.33	-	-
	ASR	0.21	0.18	-	-
ESTER	Manual	-	0.45	0.28	0.26
	ASR (A)	-	0.41	0.26	0.25
	ASR (B)	-	0.25	0.21	0.21
	ASR (C)	-	0.21	0.21	0.20
EPPS-EN	Manual	-	0.34	0.36	0.36
	ASR (A)	-	0.30	0.27	0.26
	ASR (B)	-	0.20	0.25	0.25
	ASR (C)	-	0.19	0.23	0.24
EPPS-ES	Manual	-	0.31	0.28	0.28
	ASR (A)	-	0.24	0.29	0.29
	ASR (B)	-	0.20	0.27	0.25
	ASR (C)	-	0.23	0.23	0.22

Table 3. Overview of past QAST results (best accuracy scores).

Generally speaking, a loss in accuracy is observed when dealing with automatic transcriptions instead of manual transcriptions. This difference is larger for tasks where the ASR word error rate is higher. Another observation concerns the loss of accuracy when dealing with different word error rates. Generally speaking, higher WER results in lower accuracy. Nonetheless, the results indicate that if a QA system performs well on manual transcriptions it also performs reasonably well on high quality automatic transcriptions.

The 2008 data sets were re-used in QAST 2009, where a new question creation method has been set up to generate spontaneous spoken questions. The overall absolute results were worse compared to 2008; which points to a globally harder task. The question development method produces requests which qualitatively seem to be more different to what is found in the documents compared to questions built after reading the documents (as in 2007 and 2008). In our opinion, that method, while leading to a harder problem, puts the task closer to a real, usable application.

The detailed results of the QAST campaigns can be found in the working notes of the CLEF 2007 [2], CLEF 2008 [3] and CLEF 2009 [4] workshops.

5. Evaluation Package

The QAST evaluation data and tools will be made publicly available to the research community as part of the "QAST 2007-2009 Evaluation Package" which will be distributed by ELDA through the ELRA catalogue[3].

The complete evaluation package contains all the necessary resources to enable any developer to benchmark his systems and compare results to those obtained during the official evaluation. The QAST Evaluation Package consists of the following:

- Description of the content of the package, and of the QAST evaluations (tasks, data, metrics, etc.),
- All data sets (corpora and question sets),
- Participants' submissions and results,
- Scoring tools.

The QAST Evaluation Package will be released as part of the CLEF Evaluation Packages published in 2010.

6. Conclusion

This paper has given an overview of the evaluation protocol and tools that were developed for the CLEF-QAST evaluation campaigns. In particular, it introduces a methodology for a semi-automatic evaluation of QAST systems based on time slot comparisons. These tools and methods will be further developed in next QAST evaluation campaigns.

The QAST 2007-2009 evaluation package is publicly available to the community through the ELDA Catalog. Its goal is to enable external players to benchmark their system and compare their results with those obtained during the official evaluation campaign. It will be distributed through the ELRA catalogue.

7. Acknowledgements

This work has been partially supported by the European project Treble-CLEF (ICT-1-4-1 215231).

The work of UPC (Universitat Politècnica de Catalunya) as coordinator of the QAST tracks was partially supported by the KNOW2 project (TIN2009-14715-C04-04).

The work of LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) as co-organizer of the QAST tracks has been partially financed by OSEO under the Quaero program.

The work of UPV (Universitat Politècnica de València) as co-organizer of the QAST 2009 track was partially supported by the TEXTENTERPRISE 2.0 project (TIN2009-13391-C04-03).

[3] ELRA Catalogue of Language Resources: http://catalog.elra.info/

8. References

[1] Turmo J., Surdeanu M., Galibert O., Rosset S., "Language Technologies: Question Answering in Speech Transcripts", Chapter 8 in: Alex Waibel, Rainer Stiefelhagen (Eds.), "*Computers in the Human Interaction Loop*", Springer, 2009.

[2] Turmo J., Comas P. R., Ayache C., Mostefa D., Rosset S., Lamel L., "Overview of QAST 2007", *Working Notes of CLEF 2007*, September 2007.

[3] Turmo J., Comas P. R., Rosset S., Lamel L., Moreau N., Mostefa D., "Overview of QAST 2008", *Working Notes of CLEF 2008*, September 2008.

[4] Turmo J., Comas P. R., Rosset S., Galibert O., Moreau N., Mostefa D., Rosso P. and Buscaldi D., "Overview of QAST 2009", *Working Notes of CLEF 2009*, October 2009.

[5] Lamel L., Adda G., Bilinski E. and Gauvain J.-L., "Transcribing Lectures and Seminars", *Proceedings of InterSpeech'05*, Lisbon, Portugal, 2005.

[6] Burger S., "The CHIL RT07 Evaluation Data", *Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, May 2007.

[7] Hain T., Burget L., Dines J., Garau G., Karafiat M., Lincoln M., Vepa J. and Wan V., "The AMI system for the Transcription of meetings", *Proceedings of IEEE ICASSP'07*, Hawaii, 2007.

[8] Galliano S., Geoffrois E., Gravier G., Bonastre J.F., Mostefa D. and Choukri K., "Corpus Description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News", *Proceedings of LREC'06*, Genoa, Italy, 2006.

[9] Mostefa D., Hamon O., Choukri K., "Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the First Evaluation Campaign**"**, *Proceedings of LREC'06*, Genoa, Italy, 2006.

[10] Buscaldi D., Rosso P., Turmo J. and Comas P.R, "Towards the Evaluation of Voice-Activated Question Answering Systems: Spontaneous Questions for QAST 2009", *Proceedings of the III Jornadas PLN-TIMM*, Madrid, Spain, 2009.

[11] Ayache C., Grau B., Vilnat A., "EQueR: the French Evaluation campaign of Question Answering system EQueR/EVALDA", *Proceedings of LREC'06*, Genoa, Italy, 2006.