Overview of QAST 2009

J. Turmo[1], P.R. Comas[1], S. Rosset[2], O. Galibert[2], N. Moreau[3], D. Mostefa[3], P. Rosso[4] and D. Buscaldi[4]
[1]TALP Research Centre (UPC). Barcelona. Spain
{turmo,pcomas}@lsi.upc.edu
[2]LIMSI. Paris. France
{rosset,olivier.galibert}@limsi.fr
[3]ELDA/ELRA. Paris. France
{moreau,mostefa}@elda.org
[4]NLE Lab. - ELiRF Research Group (UPV). Spain
{prosso,dbuscaldi}@dsic.upv.es

Abstract

This paper describes the experience of QAST 2009, the third time a pilot track of CLEF has been held aiming to evaluate the task of Question Answering in Speech Transcripts. Four sites submitted results for at least one of the three scenarios (European Parliament debates in English and Spanish and broadcast news in French). In order to assess the impact of potential errors of automatic speech recognition, for each task manual transcripts and three different ASR outputs were provided. In addition an original method of question creation was tried in order to get spontaneous oral questions resulting in two sets of questions (spoken and written). Each participant who had chosen a task, was asked to submit a run for each condition. The QAST 2009 evaluation framework is described, along with descriptions of the three scenarios and their associated data, the system submissions for this pilot track and the official evaluation results.

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

Question Answering (QA) technology aims at providing answers to natural language questions. Current QA technology is focused mainly on the mining of written text sources for extracting the answer to written questions from both open-domain and restricted-domain document collections [7, 3]. However, most human interaction occurs through speech, e.g. meetings, seminars, lectures, telephone conversations. All these scenarios provide large amounts of information that could be mined by QA systems. As a consequence, the exploitation of speech sources brings QA a step closer to many real world applications in which spontaneous oral questions or written questions can be involved. The QAST 2009 track aims at investigating the problem of answer spontaneous oral questions and written questions using audio documents.

Current text-based QA systems tend to use technologies that require text written in accordance with standard norms for written grammar. The syntax of speech is quite different than that of written language, with more local but less constrained relations between phrases, and punctuation, which gives boundary cues in written language, is typically absent. Speech also contains disfluencies, repetitions, restarts and corrections. Moreover, any practical application of search in speech requires the transcriptions to be produced automatically, and the Automatic Speech Recognizers (ASR)introduce a number of errors. Therefore current techniques for text-based QA need substantial adaptation in order to access the information contained in audio documents, and probably to analyse oral questions. Preliminary research on QA in speech transcriptions was addressed in QAST 2007 and QAST 2008, pilot evaluation tracks at CLEF in which systems attempted to provide answers to written factual and definitional written questions by mining speech transcripts of different scenarios [5, 6].

This paper provides an overview of the third QAST pilot evaluation. 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 reports and discusses the achieved results, followed by some conclusions in Section 6.

2 The QAST 2009 task

The aim of this third year of QAST is to provide a framework in which QA systems can be evaluated in a real scenario, where the answers of both spontaneous oral questions and written questions have to be extracted from speech transcriptions, these transcriptions being manually and automatically generated. There are five main objectives to this evaluation:

- Motivating and driving the design of novel and robust QA architectures for speech transcripts;
- Measuring the loss due to the inaccuracies in state-of-the-art ASR technology;
- Measuring this loss at different ASR performance levels given by the ASR word error rate;
- Measuring the loss when dealing with spontaneous oral questions
- Motivating the development of monolingual QA systems for languages other than English.

In the 2009 evaluation, as in the 2008 evaluation, an answer is structured as a simple [answer string, document id] pair where the answer string contains nothing more than the full and exact

Spontaneous oral question: ***When did the bombing of Fallujah eee took take place?***
Written question: ***When did the bombing of Fallujah take place?***

Manual transcript: *(%hesitation) a week ago President the American (%hesitation) occupation forces (%hesitation) m() m() m() marched into Fallujah and they (%hesitation) bombarded (%hesitation) m() murdered and have been persecuting everyone in the city .*
Answer: *a week ago*

Extracted portion of an **automatic transcript (CTM file format):**

```
(...)
20041115_1705_1735_EN_SAT 1 1081.588 0.050 a 0.9595
20041115_1705_1735_EN_SAT 1 1081.638 0.190 week 0.9744
20041115_1705_1735_EN_SAT 1 1081.828 0.350 ago 0.9743
20041115_1705_1735_EN_SAT 1 1082.338 0.630 President 0.9576
20041115_1705_1735_EN_SAT 1 1083.648 0.310 the 0.9732
20041115_1705_1735_EN_SAT 1 1084.008 0.710 American 0.9739
20041115_1705_1735_EN_SAT 1 1085.078 0.450 occupation 0.9739
20041115_1705_1735_EN_SAT 1 1085.528 0.640 forces 0.9741
20041115_1705_1735_EN_SAT 1 1086.858 1.730 and 0.9742
20041115_1705_1735_EN_SAT 1 1089.098 0.170 we 0.6274
20041115_1705_1735_EN_SAT 1 1089.308 0.480 must 0.9571
20041115_1705_1735_EN_SAT 1 1089.948 0.300 into 0.9284
20041115_1705_1735_EN_SAT 1 1090.368 0.130 for 0.3609
20041115_1705_1735_EN_SAT 1 1090.498 0.130 the 0.3609
20041115_1705_1735_EN_SAT 1 1090.698 0.240 Chair 0.2233
20041115_1705_1735_EN_SAT 1 1091.678 0.600 and 0.9755
20041115_1705_1735_EN_SAT 1 1092.798 0.400 they 0.9686
20041115_1705_1735_EN_SAT 1 1093.598 0.530 bombarded 0.8314
(...)
```

Answer: 1019.228 1019.858

Figure 1: Example query and response from manual (top) and automatic (bottom) transcripts.

answer, and the document id is the unique identifier of the document supporting the answer. For the tasks on automatic speech transcripts, the answer string consisted of the <start-time> and the <end-time> giving the position of the answer in the signal.

Figure 1 illustrates this point. Given the manually transcribed spontaneous oral question *When did the bombing of Fallujah eee took take place?* corresponding to the written question *When did the bombing of Fallujah take place?*, the figure compares the expected answer in a manual transcript (the text *a week ago*) and in an automatic transcript (the time segment *1081.588 1082.178*). Note that *Fallujah* was wrongly recongnized as *for the Chair* by the ASR. A system can provide up to 5 ranked answers per question.

A total of six tasks were defined for this third edition of QAST covering three scenarios: English questions related to European Parliament sessions in English (T1a and T1b), Spanish questions related to European Parliament sessions in Spanish (t2a and T2b) and French questions related to French Broadcast News (t3a and T3b). The complete set of tasks is:

- T1a: QA of English written questions in the manual and automatic transcriptions of European Parliament Plenary sessions in English (EPPS English corpus).

- T1b: QA of manual transcriptions of English spontaneous oral questions in the manual and automatic transcriptions of European Parliament Plenary sessions in English (EPPS English corpus).
- T2a: QA of Spanish written questions in the manual and automatic transcriptions of European Parliament Plenary sessions in Spanish (EPPS Spanish corpus).
- T2b: QA of manual transcriptions of Spanish spontaneous oral questions in the manual and automatic transcriptions of European Parliament Plenary sessions in Spanish (EPPS Spanish corpus).
- T3a: QA of French written questions in manual and automatic transcriptions of broadcast news for French (ESTER corpus)
- T3b: QA of manual transcriptions of French spontaneous oral questions in manual and automatic transcriptions of broadcast news for French (ESTER corpus)

3 Evaluation protocol

3.1 Data collections

The QAST 2009 data is derived from three different resources, each one corresponding to a different language (English, Spanish and French):

- English parliament (EPPS EN): The **TC-STAR05 EPPS English corpus** [4] contains 3 hours of recordings in English corresponding to 6 sessions of the European Parliament. The data was used to evaluated speech recognizers in the TC-STAR project. There are 3 different automatic speech recognition outputs with different word error rates (10.6%, 14% and 24.1%) . The manual transcriptions were done by ELDA.
- Spanish parliament (EPPS ES): The **TC-STAR05 EPPS Spanish corpus** [4] is comprised of three hours of recordings in Spanish corresponding to 6 sessions of the European Parliament. The data was used to evaluate Spanish ASR systems developed in the TC-STAR project. There are 3 different automatic speech recognition outputs with different word error rates (11.5%, 12.7% and 13.7%). The manual transcriptions were done by ELDA.
- French broadcast news (French BN): The test portion of the **ESTER corpus** [2] contains 10 hours of broadcast news recordings in French, comprising 18 shows from different sources (France Inter, Radio France International, Radio Classique, France Culture, Radio Television du Maroc). There are 3 different automatic speech recognition outputs with different error rates (11.0%, 23.9% and 35.4%). The manual transcriptions were produced by ELDA.

These three collections are the same than the ones used last year for the QAST 2008 evaluation campaign.

European Parliament and Broadcast News data are usually referred to as *prepared speech*. Although they typically have few interruptions and turn-taking problems when compared to actual *spontaneous speech*, many of the characteristics of spoken language are still present (hesitations, breath noises, speech errors, false starts, mispronunciations and corrections).

3.2 Questions and answer types

For each of the three languages, two sets of manually transcribed spontaneous oral questions and their respective written questions have been created and provided to the participants, the first for development purposes and the second for the evaluation:

- Development sets (released on the 25th of March 2009):
  - EPPS EN: 50 transcribed questions and their respective written questions.
  - EPPS ES: 50 transcribed questions and their respective written questions.
  - French BN: 50 transcribed questions and their respective written questions.
- Evaluation sets (released on the 1st of June 2009):
  - EPPS EN: 100 transcribed questions and their respective written questions.
  - EPPS ES: 100 transcribed questions and their respective written questions.
  - French BN: 100 transcribed questions and their respective written questions.

For each language, both the development and evaluation sets were created from the whole document collection (i.e. the 6 European Parliament sessions for English and Spanish, and the 18 Broadcast News shows for French). In other words, there was no collection split between a development data set and an evaluation data set as was done last year.

As for last year, two types of questions were considered: factual questions and definitional ones. The expected answer to a factual question is a *named entity.* There were 6 types of factual question this year, each corresponding to a particular category of named entities:

- **Person:** names of humans, real and fictional, fictional or real non-human individuals.
  Ex: *Mirjam Killer*, *John*, *Jesus*, etc.
- **Organisation:** names of business, multinational organizations, political parties, religious groups, etc.
  Ex: *CIA*, *IBM*, but also named entities like *Washington* when they display the characteristics of an organisation.
- **Location:** geographical, political or astronomical entities.
  Ex: *California*, *South of California*, *Earth*, etc.
- **Time:** a date or a specific moment in time, absolute and relative time expressions.
  Ex: *March 28th*, *last week*, *at four oclock in the morning*, etc.
- **Measure:** measures of length, width or weight, etc. Generally, a quantity and a unit of measurement.
  Ex: *five kilometers*, *20 hertz*, etc. But also ages, period of time, etc.

This is less than the 10 categories used for the 2007 and 2008 evaluations. Some categories have not been considered this year because no occurence were found in the collected set of sponteaneous questions (*Color*, *Shape*, *Language*, *System*, *Material*).

The definition questions are questions such as *What is the CDU?* and the answer can be anything. In this example, the answer would be *political group.* This year, the definition questions are subdivided into three types:

- **Person:** question about someone.
  Q: *Who is George Bush?*
  R: *The President of the United States of America.*

- **Organisation:** question about an organisation.
  Q: *What is Cortes?*
  R: *Parliament of Spain.*

- **Other:** questions about technology, natural phenomena, etc.
  Q: *What is the name of the system created by AT&T?*
  R: *The How can I help you system.*

For each language a number of 'NIL' questions (i.e., questions having no answer in the document collection) have been selected. The distribution of the different types of questions across the three collections is shown in Table 1.

Type	Factual	Definition	NIL
T1 (English)	75%	25%	18%
T2 (Spanish)	55%	45%	23%
T3 (French)	68%	32%	21%

Table 1: Distribution of question types per task: T1 (EPPS EN), T2 (EPPS ES), T3 (French BN).

The question sets are formatted as plain text files, with one question per line (see the QAST 2008 Guidelines[1]). The procedure to generate the questions is described in the following section.

3.2.1 Question generation

A novel feature in QAST 2009 was the introduction of spontaneous oral questions. The main issue in the generation of this kind of questions was how to obtain spontaneity. The solution adopted was to set up the following procedure for question generation:

1. Passage generation: a set of passages was randomly extracted from the document collection. A single passage was composed by the complete sentences included in a text window of 720 characters.

2. Question generation: human question generators were randomly assigned a number of passages (varying from 2 to 4). They had to read each passage and then to formulate one or more questions based on the passage they just read about information not present in it.

3. Question transcription: precise transcriptions of the oral spontaneous questions were made, including hesitations, etc.
   Ex: *(%hesitation) What (%hesitation) house is the pres() the president elect being elected to?*

4. Question filtering: some questions were filtered out from the set of generated questions because their answer types were not allowed or because they did not have answer in the document collection. The resulting questions were usable questions.

[1] http://www.lsi.upc.edu/~qast: News

5. Written question generation: the usable questions were re-written by removing speech disfluencies, correcting the syntax and simplifying the sentence when necessary.
   Ex: *What house does the president run?*

6. Question selection: the final set of development questions and test questions were selected by ELDA from the usable questions.

The allowed question types were the following:

- *definition*: person, organisation, object and other
- *factoid*: person, location, organisation, time (includes date), measure and language

However, the types "language" for factual questions and "object" for definition questions did not occur among the generated questions.

A preliminary evaluation of the generated questions was carried out in order to determine how many usable questions could be produced by a human reader. The results of this evaluation show that the percentage of usable questions produced by the questions generator was between 47% and 58% of the total questions produced, depending on the speakers knowledge of the task guidelines. These figures show that the produced questions were more than the number of questions actually presented to participants in QAST 2009. Most unusable questions were due to the fact that human question generators *forgot* the guidelines many times while asking their questions. Table 3.2.1 shows the number of questions recorded, the resulting usable questions and the average of the length in words per question for each language.

	#speaker	#questions recorded	#usable questions	avg. #words
English	12	1096	616	9.1
French	7	485	335	7.7
Spanish	11	403	313	7.1

Table 2: Details of the questions generated for each language.

3.3 Human judgment

As in 2008, the answer files submitted by participants have been manually judged by native speaking assessors, who considered the correctness and exactness of the returned answers. They also checked that the document labeled with the returned document ID supports the given answer. One assessor evaluated the results, and another assessor manually checked each judgment of the first one. Any doubts about an answer was solved through various discussions. The assessors used the QASTLE[2] evaluation tool developed in Perl (at ELDA) to evaluate the systems' results. A simple window-based interface permits easy, simultaneous access to the question, the answer and the document associated with the answer.

After each judgment the submission files were modified by the interface, adding a new element in the first column: the answer's evaluation (or judgment). The four possible judgments (also used at TREC[7]) correspond to a number ranging between 0 and 3:

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

- 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.
- 2 inexact: the answer-string contains a correct answer and the docid supports it, but the string has bits of the answer missing or contains additional texts (longer than it should be).
- 3 unsupported: the answer-string contains a correct answer, but is not supported by the docid.

3.4 Measures

The two following metrics (also used in CLEF) were used in the QAST evaluation:

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

4 Submitted runs

A total of four groups from four different countries submitted results for one or more of the proposed QAST 2009 tasks. Due to various reasons (technical, financial, etc.), eight other groups registered but were not be able to submit any results.

The four participating groups were:

- INAOE, Instituto Nacional de Astrofísica, Optica y Electríca, Mexico;
- LIMSI, Laboratoire d'Informatique et de Mécanique des Sciences de l'Ingénieur, France;
- TOK, Tokyo Institute of Technology, Japan;
- UPC, Universitat Politècnica de Catalunya, Spain.

All groups participated to task T1 (EPPS EN), UPC and LIMSI participated to task T2 (EPPS ES) and only LIMSI dealt with task T3 (French BN). Each participant could submit up to 48 submissions (2 runs per task and transcription). In order to allow comparisons on the performance of the systems when using different WER levels in the transcriptions, it was mandatory for each task to submit results for all the data: the manual transcriptions and the three ASR outputs (automatic transcriptions).

Table 3 shows the number of submitted runs per participant and task. The number of submissions ranged from 8 to 32. The characteristics of the systems used in the submissions are summarized in Table 4. A total of 86 submissions were evaluated with the distribution across tasks shown in the bottom row of the table.

Participant	T1a	T1b	T2a	T2b	T3a	T3b
INAOE	8	8	-	-	-	-
LIMSI	5	5	5	5	5	5
TOK	4	4	-	-	-	-
UPC	8	8	8	8	-	-
Total	25	25	13	13	5	5

Table 3: Submitted runs per participant and task.

System	Enrichment	Question classification	Doc./Passage Retrieval	Factual Answer Extraction	Def. Answer Extraction	NERC
inaoe1	words and NEs	hand-crafted rules	Indri	passage selection based on NEs of the question type	-	regular expressions
inaoe2	same plus phonetics					
limsi1	words, lemmas, morphologic derivations, synonymic relations and extended NEs	hand-crafted rules	passage ranking based on search descriptors	ranking based on distance and redundancy	specific index for known acronyms	hand-crafted rules with statistical POS
limsi2				ranking based on bayesian modelling		
tok1	words and word classes derived from training data - question-answer pairs	-	sentence ranking based on statistical models	ranking based on analogy between input question and question in the training data	-	-
upc1	words, NEs lemmas and POS	perceptrons	passage ranking through iterative query relaxation	ranking based on keyword distance and density	-	hand-crafted rules, gazeetters and perceptrons
upc2	same plus phonetics		addition of approximated phonetic matching			

Table 4: Characteristics of the systems that participated in QAST 2009.

5 Results

The results for the three tasks in manual transcribed data are presented in Tables 5 to 7, according to the question types (factual, definitional and all questions).

System	Questions	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
INAOE1	Written	44	0.38	26.7%	10	0.31	28.0%	0.36	27%
	Spoken	28	0.27	21.3%	7	0.26	24.0%	0.27	22%
INAOE2	Written	42	0.38	28.0%	9	0.30	28.0%	0.36	**28%**
	Spoken	38	0.35	25.3%	9	0.30	28.0%	0.34	26%
LIMSI1	Written	42	0.39	29.3%	11	0.28	20.0%	0.36	27%
	Spoken	39	0.36	25.3%	10	0.24	16%	0.33	23%
LIMSI2	Written	32	0.31	22.7%	13	0.36	24.0%	0.32	23%
	Spoken	30	0.26	18.7%	11	0.30	20.0%	0.27	19%
TOK1	Written	11	0.10	6.7%	3	0.03	0.0%	0.08	5%
	Spoken	11	0.08	4.0%	3	0.03	0.0%	0.06	3%
UPC1	Written	32	0.27	18.7%	8	0.29	28.0%	0.28	21%
	Spoken	19	0.15	9.3%	2	0.05	4.0%	0.12	8%
UPC2	Written	35	0.31	22.7%	8	0.29	28.0%	0.31	24%
	Spoken	18	0.15	9.3%	2	0.05	4.0%	0.12	8%

Table 5: Results for task T1, English EPPS, manual transcripts (75 factual questions and 25 definitional ones).

System	Questions	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
LIMSI1	Written	32	0.56	45.5%	29	0.36	28.6%	0.45	**36.0%**
	Spoken	32	0.56	45.5%	30	0.37	28.6%	0.45	36.0%
LIMSI2	Written	26	0.41	29.5%	23	0.28	19.6%	0.34	24.0%
	Spoken	26	0.41	29.5%	23	0.28	19.6%	0.34	24.0%
UPC1	Written	16	0.24	15.9%	10	0.16	14.3%	0.20	15.0%
	Spoken	20	0.34	27.3%	9	0.13	10.7%	0.22	18.0%
UPC2	Written	20	0.29	18.2%	10	0.14	10.7%	0.20	14.0%
	Spoken	20	0.33	27.3%	9	0.13	8.9%	0.22	17.0%

Table 6: Results for task T2, Spanish EPPS, manual transcripts (44 factual questions and 56 definitional ones).

System	Questions	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
LIMSI1	Written	38	0.35	23.5%	22	0.47	37.5%	0.39	**28.0%**
	Spoken	39	0.36	23.5%	20	0.46	37.5%	0.39	28.0%
LIMSI2	Written	38	0.34	22.1%	22	0.47	37.5%	0.38	27.0%
	Spoken	39	0.36	23.5%	20	0.46	37.5%	0.39	28.0%

Table 7: Results for task T3, French Broadcast News, manual transcripts (68 factual questions and 32 definitional ones).

The results for the three tasks in automatically transcribed data are presented in Tables 8 to 10, according to the question types (factual, definitional and all questions).

7 systems participated in the T1 (English) task on manual transcripts and 6 on automatic transcripts.

On manual transcripts, the accuracy ranged from 28% to 5% (for written questions) and from 26% to 3% (for spoken questions).

For five of the systems, we observe a relatively small difference between written and spoken questions (from 2% to 5% loss going from written questions to spoken questions). The other two systems encountered a significant loss (13% and 16% of difference between written and spoken questions).

There were three approaches for QA on automatic speech transcripts used by the systems. The LIMSI and UPC on all ASRs and INAOE on ASR_A and ASR_B took the ASR output at the only available information. INAOE on ASR_C used information extracted from all the ASR outputs, keeping ASR_C as primary. This approach could represent an application where multiple ASR outputs from different systems are available. Combining outputs from varied systems is a standard method in speech recognition to obtain a better word error rate [1], it is interesting to see if the same kind of method can be used at a more semantic level. The TOK system on the other hand used sentence segmentation information from the manual transcripts and applied it to the automatic transcripts. While such a segmentation information is not available in the transcriptions given, ASR systems do generate an acoustically motivated segmentation as a step of their processing. The TOK approach could then be considered as using an optimistic approximation of this automatically generated segmentation information. In any case, comparing systems and estimating the impact of WER can only be done on "pure" systems (LIMSI and UPC on all ASRs and INAOE on ASR_A and ASR_B).

On the ASR transcripts for the pure systems, the accuracy ranged for the best ASR (10.6%

ASR	System	Questions	Factual			Definitional			All	
			#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
ASR_A 10.6%	INAOE1	Written	35	0.32	24.0%	6	0.21	20.0%	0.30	23.0%
		Spoken	34	0.33	25.3%	6	0.21	20.0%	0.30	24.0%
	INAOE2	Written	35	0.32	22.7%	7	0.22	20.0%	0.29	22.0%
		Spoken	34	0.32	24.0%	7	0.22	20.0%	0.29	23.0%
	LIMSI1	Written	32	0.34	28.0%	10	0.25	20.0%	0.31	**26.0%**
		Spoken	30	0.31	25.3%	11	0.29	24.0%	0.30	25.0%
	TOK1	Written	13	0.08	4.0%	3	0.04	0.0%	0.07	3.0%
		Spoken	12	0.07	2.7%	4	0.08	4.0%	0.07	3.0%
	UPC1	Written	29	0.27	18.7%	7	0.26	24.0%	0.27	20.0%
		Spoken	11	0.08	5.3%	2	0.06	4.0%	0.08	5.0%
	UPC2	Written	30	0.26	18.7%	6	0.24	24.0%	0.26	20.0%
		Spoken	12	0.09	5.3%	1	0.04	4.0%	0.08	5.0%
ASR_B 14.0%	INAOE1	Written	23	0.22	16.0%	6	0.21	20.0%	0.22	17.0%
		Spoken	23	0.21	13.3%	7	0.25	24.0%	0.22	16.0%
	INAOE2	Written	24	0.22	16.0%	6	0.21	20.0%	0.22	17.0%
		Spoken	24	0.21	13.3%	7	0.25	24.0%	0.22	16.0%
	LIMSI1	Written	24	0.27	22.7%	8	0.20	16.0%	0.25	**21.0%**
		Spoken	24	0.26	21.3%	9	0.24	20.0%	0.25	21.0%
	TOK1	Written	9	0.06	4.0%	3	0.03	0.0%	0.06	3.0%
		Spoken	10	0.06	2.7%	3	0.06	4.0%	0.06	3.0%
	UPC1	Written	26	0.24	17.3%	7	0.26	24.0%	0.24	19.0%
		Spoken	11	0.08	4.0%	2	0.06	4.0%	0.08	4.0%
	UPC2	Written	29	0.26	20.0%	7	0.25	24.0%	0.26	21.0%
		Spoken	12	0.08	4.0%	2	0.05	4.0%	0.07	4.0%
ASR_C 24.1%	INAOE1	Written	29	0.31	26.7%	5	0.20	20.0%	0.28	**25.0%**
		Spoken	28	0.30	26.7%	5	0.20	20.0%	0.28	25.0%
	INAOE2	Written	29	0.30	25.3%	6	0.21	20.0%	0.28	24.0%
		Spoken	28	0.29	24.0%	6	0.21	20.0%	0.27	23.0%
	LIMSI1	Written	23	0.26	24.0%	8	0.19	12.0%	0.24	21.0%
		Spoken	24	0.24	21.3%	9	0.23	16.0%	0.24	20.0%
	TOK1	Written	17	0.12	5.3%	5	0.08	4.0%	0.11	5.0%
		Spoken	19	0.11	4.0%	5	0.12	8.0%	0.11	5.0%
	UPC1	Written	22	0.21	16.0%	6	0.24	24.0%	0.22	18.0%
		Spoken	10	0.08	5.3%	1	0.04	4.0%	0.07	5.0%
	UPC2	Written	26	0.24	17.3%	6	0.24	24.0%	0.24	19.0%
		Spoken	11	0.08	4.0%	1	0.04	4.0%	0.07	4.0%

Table 8: Results for task T1, English EPPS, automatic transcripts (75 factual questions and 25 definitional ones).

of WER) from 26% (written questions) to 5% (spoken questions). Accuracy goes down with increased word error rate giving a roughly 5% loss for ASR_B and ASR_C compared to ASR_A. It is interesting to note that the differences between ASR_B (WER 14%) and ASR_C (WER 24%) are negligible. The INAOE multi-ASR approach paid off by giving an overall result better than what was obtained by the same system on the best ASR only.

We notice that the impact of written vs spoken questions is similar than for manual transcriptions, with two systems taking an heavy loss and the others not showing a significant difference.

Four systems (2 from LIMSI and 2 from UPC) participated in the T2 (Spanish) task on manual transcripts and 3 systems (1 from LIMSI and 2 from UPC) on automatic transcripts.

On manual transcripts, the accuracy ranged from 36% (written questions and spoken questions) to 14% (written questions) and 17% (spoken questions). The differences between written questions and spoken questions is very low (from 0% to 3%). The same kind of behaviour is observed on the automatic transcripts tasks, with a loss due to the speech recognition errors and no significant difference between written and spoken questions.

ASR	System	Questions	Factual			Definitional			All	
			#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
ASR_A 11.5%	LIMSI1	Written	20	0.37	31.8%	22	0.29	23.2%	0.32	**27.0%**
		Spoken	20	0.37	31.8%	21	0.27	21.4%	0.31	26.0%
	UPC1	Written	8	0.15	13.6%	2	0.01	0.0%	0.07	6.0%
		Spoken	6	0.14	13.6%	2	0.01	0.0%	0.07	6.0%
	UPC2	Written	12	0.20	18.2%	3	0.02	0.0%	0.10	8.0%
		Spoken	12	0.24	22.7%	3	0.03	1.8%	0.12	11.0%
ASR_B 12.7%	LIMSI1	Written	18	0.32	27.3%	19	0.26	23.2%	0.29	**25.0%**
		Spoken	18	0.32	27.3%	19	0.26	23.2%	0.29	25.0%
	UPC1	Written	12	0.18	13.6%	2	0.04	3.6%	0.10	8.0%
		Spoken	12	0.20	15.9%	1	0.02	1.8%	0.10	8.0%
	UPC2	Written	13	0.20	15.9%	3	0.02	0.0%	0.10	7.0%
		Spoken	12	0.20	15.9%	1	0.01	0.0%	0.09	7.0%
ASR_C 13.7%	LIMSI1	Written	18	0.33	29.5%	19	0.24	17.9%	0.28	23.0%
		Spoken	18	0.33	29.5%	19	0.25	19.6%	0.28	**24.0%**
	UPC1	Written	12	0.22	20.5%	4	0.05	3.6%	0.13	11.0%
		Spoken	8	0.13	11.4%	2	0.03	1.8%	0.07	6.0%
	UPC2	Written	11	0.20	18.2%	4	0.03	1.8%	0.11	9.0%
		Spoken	10	0.21	20.5%	3	0.02	0.0%	0.10	9.0%

Table 9: Results for task T2, Spanish EPPS, automatic transcripts (44 factual questions and 56 definitional ones).

ASR	System	Questions	Factual			Definitional			All	
			#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
ASR_A 11.0%	LIMSI1	Written	33	0.33	25.0%	19	0.47	37.5%	0.37	29.0%
		Spoken	32	0.33	25.0%	18	0.45	37.5%	0.37	29.0%
ASR_B 23.9%	LIMSI1	Written	25	0.29	25.0%	15	0.38	31.3%	0.32	27.0%
		Spoken	25	0.27	22.1%	13	0.35	31.3%	0.30	25.0%
ASR_C 35.4%	LIMSI1	Written	25	0.26	20.6%	13	0.33	28.1%	0.28	23.0%
		Spoken	24	0.25	19.1%	11	0.31	28.1%	0.27	22.0%

Table 10: Results for task T3, French Broadcast News, manual transcripts (68 factual questions and 32 definitional ones).

Only 2 systems (both from LIMSI) participated in the T3 (French) task on manual transcripts and one (from LIMSI) on automatic transcripts.

On manual transcripts, the accuracy ranged from from 28% (both written and spoken questions) to 27% (written questions). There is no significant differences between spoken and written questions (0% to 1% loss). The results for automatic transcriptions show very little loss compared to the manual transcriptions except for the worst ASR.

The overall absolute results were worse this year compared to last year 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. In our opinion that method, while giving an harder problem, puts us closer to a real, usable application.

6 Conclusions

In this paper, the QAST 2009 evaluation has been described. Four groups participated in this track with a total of 86 submitted runs across 3 main tasks that included dealing with different languages (English, Spanish and French), different word error rates for automatic transcriptions (from 10.5% to 35.4%) and different question types (written and spoken questions). An original question creation method has been tried succesfully to generate spontaneous spoken questions. Qualitatively, the questions were harder and more different to the formulations found in the documents compared to those produced by the traditional method of consulting the documents first. The method used this year gives an harder problem but we think that it is a more realistic one, putting us closer to a real, usable application.

Acknowledgments

This work has been jointly funded by the Spanish Ministry of Science (TEXTMESS project - TIN2006-15265-C06) and OSEO under the Quaero program. We thank to Lori Lamel, Erik Bilinski, Manuel González and Pere Vilarrubia their help to the organisation and data generation.

References

[1] J. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover). In *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–352, Santa Barbara, CA, 1997.

[2] S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of LREC'06*, pages 315–320, Genoa, 2006.

[3] 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.

[4] TC-Star. http://www.tc-star.org, 2004-2008.

[5] J. Turmo, P.R. Comas, C. Ayache, D. Mostefa, S. Rosset, and L. Lamel. Overview of qast 2007. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Pes, V. Petras, and D. Santos, editors, *8th workshop of the Cross Language Evaluation Forum (CLEF 2007). Revised Selected Papers.*, pages 249–256. LNCS, 2008.

[6] J. Turmo, P.R. Comas, S. Rosset, L. Lamel, N. Moreau, and D. Mostefa. Overview of qast 2008. In *9th workshop of the Cross Language Evaluation Forum (CLEF 2008). Revised Selected Papers. (to appear)*. LNCS, 2009.

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