Overview of QAST 2008

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

Abstract

This paper describes the experience of QAST 2008, the second time a pilot track of CLEF has been held aiming to evaluate the task of Question Answering in Speech Transcripts. Five sites submitted results for at least one of the five scenarios (lectures in English, meetings in English, broadcast news in French and European Parliament debates in English and Spanish). In order to assess the impact of potential errors of automatic speech recognition, for each task contrastive conditions are with manual and automatically produced transcripts. The QAST 2008 evaluation framework is described, along with descriptions of the five 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 relevant answers to natural language questions. Most Question Answering research has focused on mining document collections con-

taining written texts to answer written questions [3, 6]. Documents can be either open domain (newspapers, newswire, Wikipedia...) or restricted domain (biomedical papers...) but share, in general, a decent writing quality, at least grammar-wise. In addition to written sources, a lot (and growing amount) of potentially interesting information appears in spoken documents, such as broadcast news, speeches, seminars, meetings or telephone conversations. The QAST track aims at investigating the problem of question answering in such audio documents.

Current text-based QA systems tend to use technologies that require texts to have been 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. Preliminary research on QA in speech transcriptions was addressed in QAST 2007, a pilot evaluation track at CLEF 2007 in which systems attempted to provide answers to written factual questions by mining speech transcripts of seminars and meetings [5].

This paper provides an overview of the second QAST pilot evaluation. Section 2 describes the principles of this evaluation track. Sections 3 present the evaluation framework and section 4 the systems that participated. Section 5 reports and discusses the achieved results, followed by some conclusions in Section 6.

2 The QAST 2008 task

The objective of this pilot track is to develop a framework in which QA systems can be evaluated when the answers have to be found in speech transcripts, these transcripts being either produced manually or automatically. 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;
- Comparing the performance of QA systems on different kinds of speech data (prepared speech such as broadcast news (BN) or parliamentary hearings vs. spontaneous in meeting for instance);
- Motivating the development of monolingual QA systems for languages other than English.

In the 2008 evaluation, as in the 2007 pilot 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 answer, and the document id is the unique identifier of the document supporting the answer. In 2008, 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 comparing the expected answer to the question *What is the Vlaams Blok?* in a manual transcript (the text *criminal organisation*) and in an automatic transcript (the time segment *1019.228 1019.858*). A system can provide up to 5 ranked answers per question.

Question: *What is the Vlaams Blok?*

Manual transcript: *the Belgian Supreme Court has upheld a previous ruling that declares the Vlaams Blok a criminal organization and effectively bans it .*
Answer: *criminal organisation*

Extracted portion of an **automatic transcript (CTM file format):**
(...)
20041115_1705_1735_EN_SAT 1 1018.408 0.440 Vlaams 0.9779
20041115_1705_1735_EN_SAT 1 1018.848 0.300 Blok 0.8305
20041115_1705_1735_EN_SAT 1 1019.168 0.060 a 0.4176
20041115_1705_1735_EN_SAT 1 **1019.228** 0.470 criminal 0.9131
20041115_1705_1735_EN_SAT 1 **1019.858** 0.840 organisation 0.5847
20041115_1705_1735_EN_SAT 1 1020.938 0.100 and 0.9747
(...)
Answer: 1019.228 1019.858

Figure 1: Example query *What is the Vlaams Blok?* and response from manual (top) and automatic (bottom) transcripts.

A total of ten tasks were defined for this second edition of QAST covering five main task scenarios and three languages: lectures in English about *speech and language processing* (T1), meetings in English about *design of television remote controls* (T2), French broadcast news (T3) and European Parliament debates in English (T4) and Spanish (T5). The complete set of tasks are:

- T1a: QA in manual transcriptions of lectures in English.
- T1b: QA in automatic transcriptions of lectures in English.
- T2a: QA in manual transcriptions of meetings in English.
- T2b: QA in automatic transcriptions of meetings in English.
- T3a: QA in manual transcriptions of broadcast news for French.
- T3b: QA in automatic transcriptions of broadcast news for French.
- T4a: QA in manual transcriptions of European Parliament Plenary sessions in English.
- T4b: QA in automatic transcriptions of European Parliament Plenary sessions in English.
- T5a: QA in manual transcriptions of European Parliament Plenary sessions in Spanish.
- T5b: QA in automatic transcriptions of European Parliament Plenary sessions in Spanish.

3 Evaluation protocol

3.1 Data collections

The data for this second edition of QAST is derived from five different resources, covering spontaneous speech, semi-spontaneous speech and prepared speech: The first two are the same as were used in QAST 2007 [5].

- The **CHIL corpus**[1] (as used for QAST 2007): The corpus contains about 25 hours of speech, mostly spoken by non native speakers of English, with an estimated ASR Word Error Rate (WER) of 20%.
- The **AMI corpus**[2] (as used for QAST 2007): This corpus contains about 100 hours of speech, with an ASR WER of about 38%.
- French broadcast news: The test portion of the **ESTER corpus** [1] contains 10 hours of broadcast news in French, recorded 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 (WER = 11.0%, 23.9% and 35.4%). The manual transcriptions were produced by ELDA.
- Spanish parliament: The **TC-STAR05 EPPS Spanish corpus** [4] is comprised of three hours of recordings from the European Parliament in Spanish. The data was used to evaluate recognition 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.
- English parliament: The **TC-STAR05 EPPS English corpus** [4] contains 3 hours of recordings from the European Parliament in English. 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.

The spoken data cover a broader range of types, both in terms of content and in speaking style. The Broadcast News and European Parliament date are less spontaneous than the lecture and meeting speech as they are typically prepared in advance and are closer in structure to written texts. While meetings and lectures are representative of *spontaneous speech*, Broadcast News and European Parliament sessions are usually referred to as *prepared speech*. Although they typically have few interruptions and turn-taking problems when compared to meeting data, many of the characteristics of spoken language are still present (hesitations, breath noises, speech errors, false starts, mispronunciations and corrections). One of the reasons for including the additional types of data was to be closer to the textual data used to assess written QA, and to benefit from the availability of multiple speech recognizers that have been developed for these languages and tasks in the context of European or national projects [2, 1, 4].

3.1.1 Questions and answer types

For each of the five scenarios, two sets of questions have been provided to the participants, the first for development purposes and the second for the evaluation.

- Development set (11 March 2008) :
  - Lectures: 10 seminars and 50 questions.
  - Meetings: 50 meetings and 50 questions.
  - French broadcast news: 6 shows and 50 questions.
  - English EPPS: 2 sessions and 50 questions.
  - Spanish EPPS: 2 sessions and 50 questions.

[1] http://chil.server.de
[2] http://www.amiproject.org

- Evaluation set (15 June 2008):
  - Lectures: 15 seminars and 100 questions.
  - Meetings: 120 meetings and 100 questions.
  - French broadcast news: 12 shows and 100 questions.
  - English EPPS: 4 sessions and 100 questions.
  - Spanish EPPS: 4 sessions and 100 questions.

Two types of questions were considered this year: factual questions and definitional ones. For each corpus (CHIL, AMI, ESTER, EPPS EN, EPPS ES) roughly 70% of the questions are factual, 20% are definitional, and 10% are NIL (i.e., questions having no answer in the document collection).

The question sets are formatted as plain text files, with one question per line (see the QAST 2008 Guidelines[3]). The factual questions similar to those used in the 2007 evaluation. The expected answer to these questions is a Named Entity (person, location, organization, language, system, method, measure, time, color, shape and material). The definition questions are questions such as *What is the Vlaams Blok?* and the answer can be anything. In this example, the answer would be *a criminal organization*. The definition questions are subdivided into the following 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.*
- **Object:** question about any kind of objects
  Q: *What is F-15?*
  R: *combat aircraft.*
- **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.*

3.2 Human judgment

As in QAST 2007, 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 docid 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[4] evaluation tool developed in Perl (at ELDA) to evaluate the responses. A simple window-based interface permits easy, simultaneous access to the question, the answer and the document associated with the answer.

For T1b, T2b, T3b, T4b and T5b (QA on automatic transcripts) the manual transcriptions were aligned to the automatic ASR outputs to find associate times with the answers in the automatic

[3] http://www.lsi.upc.edu/~qast: News
[4] http://www.elda.org/qastle/

transcripts. The alignments between the automatic and the manual transcription were done using time information. Unfortunately, for some documents time information were not available and only word alignments were used.

After each judgment the submission files were modified, adding a new element in the first column: the answer's evaluation (or judgment). The four possible judgments (also used at TREC[6]) 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.
- 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.

3.3 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 five groups from four different countries submitted results for one or more of the proposed QAST 2008 tasks. Due to various reasons (technical, financial, etc.), three other groups registered but were not be able to submit any results.

The five participating groups were:

- CUT, Chemnitz University of Technology, Germany;
- INAOE, Instituto Nacional de Astrofica, Optica y Electrica, Mexico;
- LIMSI, Laboratoire d'Informatique et de Mécanique des Sciences de l'Ingénieur, France;
- UA, Universidad de Alicante, Spain;
- UPC, Universitat Politècnica de Catalunya, Spain.

All groups participated to task T4 (English EPPS). Only LIMSI participated to task T3 (French broadcast news). Table 1 shows the number of submitted runs per participant and task. Each

participant could submit up to 32 submissions (2 runs per task and transcription). The number of submissions ranged from 2 to 20. The characteristics of the systems used in the submissions are summarized in Table 2. A total of 49 submissions were evaluated with the distribution across tasks shown in the bottom row of Table 2.

Participant	T1a	T1b	T2A	T2b	T3a	T3b	T4a	T4b	T5a	T5b
CUT	2	-	-	-	-	-	2	-	-	-
INAOE	-	-	-	-	-	-	1	2	-	-
LIMSI	1	1	1	1	2	3	1	3	2	3
UA	-	-	-	-	-	-	1	3	-	-
UPC	1	2	1	2	-	-	1	6	1	6
Total	4	3	2	3	2	3	6	14	3	9

Table 1: Submitted runs per participant and task. T1 (English lectures), T2 (English meetings), T3 (French BN), T4 (English EPPS), T5 (Spanish EPPS).

System	Enrichment	Question classification	Doc./Passage Retrieval	Factual Answer Extraction	Def. Answer Extraction	NERC
cut1	words, NEs and POS	hand-crafted rules	pass. ranking	hand-crafted rules with fallback str. in 1st pass.	hand-crafted fallback strategy	Stanford NER, rules with classification
cut2			based on RSV	same in top-3 pass.		
inaoe1	words and NEs	hand-crafted rules	Lemur	candidate selection based on NEs	-	regular expressions
inaoe2	same plus phonetics					
limsi1	words, lemmas, morphologic derivations,	hand-crafted rules	ranking based on search descriptors	ranking based on distance and redundancy	specific index for known acronyms	hand-crafted rules with stochastic POS
limsi2	synonyms and extended NEs			tree-rewriting based distance		
ua1	words, NEs POS and n-grams	hand-crafted rules	ranking based on n-grams	ranking based on keyword distance and mutual information	-	hand-crafted rules
upc1	words, NEs lemmas and POS	perceptrons	ranking based on 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 2: Characteristics of the systems that participated in QAST 2008.

5 Results

The results for the ten QAST 2008 tasks are presented in Tables 3 to 12, according to factual questions, definitional questions, and all questions.

For manual transcriptions, the accuracy ranges from 45% (LIMSI1 on task T3a) down to 7% (UPC1 on task T5a). For automatic transcriptions, the accuracy goes from 41% (LIMSI1 on task T3b and ASR a) to 2% (UPC1 on task T5b and ASR c). Generally speaking, a loss in accuracy is observed when dealing with automatic transcriptions. Comparing the best accuracy results on manual transcription and automatic transcriptions, the loss of accuracy goes from 15% for task T2 to 4% for tasks T3 and T4 tasks. This difference is larger for tasks where the ASR word error rate is higher.

System	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
cut1	14	0.18	17.9	2	0.09	9.1	0.16	16.0
cut2	16	0.19	16.7	8	0.26	18.2	0.20	17.0
limsi1	48	0.53	47.4	4	0.18	18.2	0.45	41.0
upc1	39	0.44	38.5	4	0.18	18.2	0.38	34.0

Table 3: Results for task T1a, English lectures, manual transcripts (78 factual questions and 22 definitional ones).

System ASR 20%	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
limsi1	33	0.34	30.8	3	0.14	13.6	0.30	27.0
upc1	35	0.39	34.6	4	0.18	18.2	0.34	31.0
upc2	35	0.37	33.3	4	0.18	18.2	0.33	30.0

Table 4: Results for task T1b, English lectures, ASR transcripts (78 factual questions and 22 definitional ones).

System	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
limsi1	44	0.47	37.8	7	0.22	19.2	0.40	33.0
upc1	29	0.35	31.1	3	0.12	11.5	0.29	26.0

Table 5: Results for task T2a, English meetings, manual transcripts (74 factual questions and 26 definitional ones).

System ASR 38%	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
limsi1	23	0.21	16.2	6	0.18	15.4	0.20	16.0
upc1	19	0.20	17.6	5	0.19	19.2	0.20	18.0
upc2	16	0.16	10.8	6	0.23	23.1	0.18	14.0

Table 6: Results for task T2b, English meetings, ASR transcripts (74 factual questions and 26 definitional ones).

System	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
limsi1	45	0.50	45.3	13	0.47	44.0	0.49	45.0
limsi2	45	0.47	41.3	13	0.46	44.0	0.47	42.0

Table 7: Results for task T3a, French BN, manual transcripts (75 factual questions and 25 definitional ones).

ASR	System	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
a 11.0%	limsi1	42	0.49	44.0	9	0.33	32.0	0.45	41.0
b 23.9%	limsi1	29	0.28	22.7	10	0.34	32.0	0.30	25.0
c 35.4%	limsi1	24	0.24	20.0	7	0.26	24.0	0.24	21.0

Table 8: Results for task T3b, French BN, ASR transcripts (75 factual questions and 25 definitional ones).

System	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
cut1	12	0.16	16.0	9	0.36	36.0	0.21	21.0
cut2	12	0.16	16.0	11	0.39	36.0	0.22	21.0
inaoe1	41	0.43	37.3	6	0.21	20.0	0.38	33.0
limsi1	44	0.43	33.3	12	0.39	32.0	0.42	33.0
ua1	32	0.30	21.3	4	0.16	16.0	0.27	20.0
upc1	38	0.44	40.0	4	0.16	16.0	0.37	34.0

Table 9: Results for task T4a, English EPPS, manual transcripts (75 factual questions and 25 definitional ones).

ASR	System	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
a 10.6%	inaoe1	32	0.37	33.3	5	0.20	20.0	0.33	30.0
	inaoe2	34	0.38	32.0	5	0.20	20.0	0.33	29.0
	limsi1	24	0.23	18.7	9	0.31	28.0	0.25	21.0
	ua1	12	0.09	4.0	4	0.16	16.0	0.10	7.0
	upc1	18	0.22	20.0	4	0.17	16.7	0.21	19.0
	upc2	16	0.16	13.3	4	0.17	16.7	0.16	14.1
b 14.0%	limsi1	22	0.21	16.0	9	0.33	32.0	0.24	20.0
	ua1	12	0.11	8.0	4	0.16	16.0	0.12	10.0
	upc1	15	0.18	16.0	4	0.16	16.0	0.17	16.0
	upc2	14	0.16	13.3	4	0.16	16.0	0.16	14.0
c 24.1%	limsi1	21	0.21	16.0	8	0.30	28.0	0.23	19.0
	ua1	9	0.10	8.0	5	0.20	20.0	0.12	11.0
	upc1	11	0.11	9.3	5	0.20	20.0	0.14	12.0
	upc2	11	0.11	8.0	4	0.16	16.0	0.12	10.0

Table 10: Results for task T4b English EPPS, ASR transcripts (75 factual questions and 25 definitional ones).

System	Factual			Definitional			All	
	#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
limsi1	29	0.32	29.3	13	0.44	36.0	0.35	31.0
limsi2	29	0.32	29.3	13	0.42	32.0	0.35	30.0
upc1	9	0.11	9.3	3	0.05	0.0	0.09	7.0

Table 11: Results for task T5a, Spanish EPPS, manual transcripts (75 factual questions and 25 definitional ones).

ASR	System	Factual			Definitional			All	
		#Correct	MRR	Acc	#Correct	MRR	Acc	MRR	Acc
a 11.5%	limsi1	20	0.25	24.0	8	0.28	24.0	0.26	24.0
	upc1	5	0.05	4.0	0	0.00	00.0	0.04	3.0
	upc2	5	0.06	5.3	2	0.08	8.0	0.07	6.0
b 12.7%	limsi1	18	0.20	17.3	9	0.28	24.0	0.22	19.0
	upc1	5	0.06	5.3	0	0.00	00.0	0.05	4.0
	upc2	5	0.06	5.3	2	0.08	8.0	0.07	6.0
c 13.7%	limsi1	20	0.24	22.7	8	0.27	24.0	0.25	23.0
	upc1	2	0.03	2.7	0	0.00	00.0	0.02	2.0
	upc2	3	0.03	2.7	1	0.04	4.0	0.04	3.0

Table 12: Results for task T5b, Spanish EPPS, ASR transcripts (75 factual questions and 25 definitional ones).

Another observation concerns the loss of accuracy when dealing with different word error rates.

Generally speaking higher WER results in lower accuracy (e.g. from 30% for T4b_A to 20% for T4b_B). Strangely enough this is not completely true for the T5b task where results for ASR_C (13.7% WER) are 4% higher than for ASR_B (12.7% WER). The WER being rather close, it is probable that ASR_C errors had a smaller impact on the named entities present in the questions.

6 Conclusions

In this paper, the QAST 2008 evaluation has been described. Five groups participated in this track with a total of 49 submitted runs, across ten tasks that included dealing with different types of speech (spontaneous or prepared), different languages (English, Spanish and French) and different word error rates for automatic transcriptions (from 10.5% to 35.4%). For the tasks where the word error rate was low enough (around 10%) the loss in accuracy compared to manual transcriptions was under 5%, suggesting that QA in such documents is potentially feasible. However, even where ASR performance is reasonably good, there remain outstanding challenges in dealing with spoken language and the earlier mentioned differences from written language. The results from the QAST evaluation indicate that if a QA system which performs well on manual transcriptions it also performs reasonably well on high quality automatic transcriptions. The performance on spoken language have not yet reached the level of those in the main QA track.

Acknowledgments

This work has been jointly funded by the Spanish Ministry of Science (TEXTMESS project) and OSEO under the Quaero program.

References

[1] 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*, Genoa, 2006.

[2] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. McTait, , and K. Choukri. The ESTER evaluation campaign of Rich Transcription of French Broadcast News. In *Proceedings of LREC'04*, Lisbon, 2004.

[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. Peas, V. Petras, and D. Santos, editors, *8th workshop of the Cross Language Evaluation Forum (CLEF 2007). Revised Selected Papers*. LNCS, 2008.

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