Question Answering on Speech Transcriptions: the QAST evaluation in CLEF

L. Lamel[1], S. Rosset[1], C. Ayache[2], D. Mostefa[2], J. Turmo[3], P. Comas[3]

LIMSI-CNRS, ELDA/ELRA, TALP Research Center (UPC)
Orsay - France, Paris - France, Barcelona - Spain
{lamel,rosset}@limsi.fr, {ayache,mostefa}@elda.org, {turmo,pcomas}@lsi.upc.edu

Abstract

This paper reports on the QAST track of CLEF aiming to evaluate Question Answering on Speech Transcriptions. Accessing information in spoken documents provides additional challenges to those of text-based QA, needing to address the characteristics of spoken language, as well as errors in the case of automatic transcriptions of spontaneous speech. The framework and results of the pilot QAst evaluation held as part of CLEF 2007 is described, illustrating some of the additional challenges posed by QA in spoken documents relative to written ones. The current plans for future multiple-language and multiple-task QAst evaluations are described.

1. Introduction

There are two main paradigms used to search for information: document retrieval and precise information retrieval. In the first approach, documents matching a user query are returned. The match is often based on some keywords that were extracted from the query, and the underlying assumption is that the topic of the documents best matching the query provide a data pool from which the user might find information that suits their need. This need can be very specific (e.g. *Who is presiding the Senate?*), or it can be topic-oriented (e.g. *I'd like information about the Senate*). The user is left to filter through the returned documents to find the desired information, which is quite appropriate for the more general topic-oriented questions, and less well-adapted to the more specific one. The second approach to search, which is better suited to the specific queries, is embodied by so-called question answering (QA) systems, which return the most probable answer given a specific question (e.g. The answer to *Who won the 2005 Tour de France?* is *Lance Armstrong.*).

In the QA and Information Retrieval domains progress has been assessed via evaluation campaigns (Ayache et al., 2006; Kando, 2006; Voorhees and Buckland, 2007; Nunzio et al., 2007; Giampiccolo et al., 2007). In the Question-Answering evaluations, the systems handle independent questions and should provide one answer to each question, extracted from textual data, for both open domain and restricted domain. Since much of human interaction is via spoken language (e.g. meetings, seminars, lectures, telephone conversations), it is interesting to explore applying QA on speech data. Accessing information in spoken language requires significant departures from traditional text-based approaches in order to deal with transcripts (manual or automatic) of spontaneous speech. Much of the QA research carried by natural language groups have typically developed techniques for written texts which are assumed to have a correct syntactic and semantic structure. Spoken data is different from textual data in various ways: it contains disfluencies, false starts, speaker corrections, truncated words. The grammatical structure of spontaneous speech is quite different than for written discourse. Moreover, spoken data can be meetings which show a complete different global structure (for instance, interaction creates run-on sentences where the distance between the first part of an utterance and the last one can be very long).

In 2007, a pilot evaluation campaign, partially sponsored by the FP6 CHIL project, was carried out under the CLEF umbrella for the evaluation of **QA** systems on **S**peech **T**ranscriptions: the QAST evaluation (Turmo et al., 2007).

The remainder of this paper is organized as follows. First the next section presents the QAst 2007 tasks, and is followed by a description of the 2007 evaluation in Section 3.. This is followed by a discussion of the results and plans for the 2008 evaluation in Section 4.. The tasks for 2008 and evaluation plans have been modified based on the pilot evaluation in order to allow better comparison between textual Question-Answering and Speech Question-Answering tasks, and to assess Question-Answering on automatic speech transcripts with different error rates (reflecting the quality of the automatic speech recognition systems).

2. The QAst 2007 Tasks

The design of the QAST tasks attempted to take into account two different viewpoints. Firstly, automatic transcripts of speech data contain recognition errors which can potentially lead to incorrectly answered questions or unanswered questions. In order to measure the loss of the QA systems due to automatic speech recognition (ASR) technology, a comparative evaluation was introduced for both manual and automatic transcripts. Secondly, dealing with speech from single speakers (monologues) is different than dealing with multi-speaker interactions (dialogues). With the aim of comparing the performance of QA systems for both monologues and dialogues, two scenarios were introduced: lectures and meetings in English from the CHIL (CHIL, 2004 2007) and AMI (AMI, 2005) projects. From the combination of these two viewpoints, QAST covered the following four tasks:

- T1: Question Answering in manual transcripts of lectures

- T2: Question Answering in automatic transcripts of lectures
- T3: Question Answering in manual trancripts of meetings
- T4: Question Answering in automatic transcripts of meetings

3. The 2007 Evaluation

3.1. Data and Methodology

The data for the QAST pilot track come from two different resources, one from the CHIL lecture scenario and the other from the AMI meeting scenario.
The CHIL corpus contains about 25 hours (around 1 hour per lecture) of both manually and automatically transcribed data, with most of the data from the primary speaker (the person presenting the lecture) and a small amount of speech from the audience (mostly questions or comments). The manual transcriptions were done by ELDA and the ASR transcriptions (Lamel et al., 2005) were produced by LIMSI (Lamel et al., 2005). In addition to the best word hypotheses, a 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) (Mostefa et al., 2007).

The AMI corpus is comprised of about 100 hours (168 meetings) of speech with both manual and automatic transcriptions. The AMI Rich Transcription 2006 ASR data has been used (Hain et al., 2007)). The domain of this meetings is *design of television remote control*. The language is European English. As for the lectures, meetings have been produced with simple tags and are formatted as plain text files (ISO-8859-1) (AMI, 2005).

For each one of the scenarios, two sets of questions have been provided to the participants. The development data set (30-January-2007) had 50 questions each for Lectures (10 seminars) and Meetings (50 meetings). For testing, the evaluation data (15-June-2007) had 100 questions each for Lectures (15 seminars) and Meetings (118 meetings). The question sets were distributed as plain text files, with one question per line. All the questions in the QAST task was ”factual” questions e.g. questions whose expected answer was a Named Entity (*person*, *location*, *organization*, *language*, *system/method*, *measure*, *time*, *colour*, *shape* and *material*) such as defined in specific Named Entity guidelines. No definitional questions were given. The two data collections (CHIL and AMI corpus) were first tagged with Named Entities (NE). Then, an English native speaker created questions for each NE tagged session. So each answer was 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 were no particular restrictions on the length of an answer-string (which is usually very short), but unnecessary pieces of information have been penalized, since the answer have been marked as non-exact. Assessors have focus mainly on the responsiveness and usefulness of the answers. A correct answer was defined in QAST as the token sequence comprised of the smallest number of tokens that are required to contain the correct answer in the audio stream, and its corresponding automatic transcript.

For example, consider 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 transcription:

> *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.*

For the question: *which organisation has worked with the University of Karlsruhe on the meeting transcription system?*, the answer found in the manual transcription is *Carnegie Mellon* whereas in the automatic transcription it is *coming around*.

The submitted files were assessed by English native speakers. 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 judgment 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 (QASTLE, 2007). A simple interface permits easy access to 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. For T2 and T4 the correct answer is defined as the minimal sequence of words that overlaps the reference answer in the manual transcript.

3.2. Results

Due to some problems (typos, answer type, etc.), 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 T1 and T2, and 96 for T3. In addition and due to missing time information at word level for some of the AMI seminars, seven questions have been deleted from the scoring results. In total, the results for T4 have been calculated on the basis of 93 questions.

Range of results	T1	T2	T3	T4
Best Acc.	0.51	0.36	0.25	0.21
Best MRR	0.53	0.37	0.31	0.22
Worst Acc.	0.05	0.02	0.16	0.06
Worst MRR	0.09	0.05	0.22	0.10

Table 1: Best and worst accuracy and MRR on the four tasks: T1, T2, T3 and T4. The results do not all come from the same system, but summarize the the best system results for the various conditions.

Table 1 summarizes the Accuracy and Mean Reciprocal Rank (MRR) obtained on the four tasks. For task T1 (lectures/manual transcript), the accuracy ranged from 0.05 to 0.51, whereas for task T2 (lectures/automatic transcript), the accuracy ranged from 0.02 to 0.36. For the meeting tasks, the accuracy for with the manual transcripts (T3) ranged from 0.16 to 0.25 and for automatic transcripts (T4) from 0.06 to 0.21. The best MRR ranges from about than 0.2 (T4) to over 0.5 (T1) across the tasks, with as expected better results on manual transcriptions than on ASR outputs.

4. Discussion

These initial results are very encouraging, demonstrating that QA technology is able to deal with spoken data. Meanwhile, the difference in the accuracy of systems applied to the manual and automatic transcripts, that is, between T1 and T2 (from 0.22 to 0.16 in average) and T3 and T4 (from 0.21 to 0.13) drops by over 36% when applied to automatic transcriptions. These observations (and others) have led to the change of several points in the 2008 evaluation campaign. One contrast will be to use multiple recognizer hypotheses with different word error rates (WER) with the objective of assessing the dependency of QA performance on WER. Another extension is to evaluate different subtasks (information retrieval for QA and answer extraction), the objective being to study which part of a QA system is more sensitive to word error rate. The evaluation will also be extended to include two additional languages (French and Spanish) and data types (European Parliament data and Broadcast news). This is a big change in the type of spoken data both in terms of content and in speaking style. The Broadcast News and European Parliament discourses 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 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 present (hesitations, breath noises, speech errors, false starts, mispronunciations and corrections) are still present. One of the reasons for including the additional types of data was to be closer to the type of 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 (Gravier et al., 2004; Galliano et al., 2006; TC-Star, 2004 2008). In order to avoid some alignment problems encountered in the 2007 evaluation on automatic speech transcripts, the automatic speech transcripts are provided with time stamps.

4.1. Data and evaluation for 2008

In the 2008 QAst track, 10 sub-tasks have been defined:

- T1a: Question Answering in manual transcriptions of lectures (CHIL corpus)
- T1b: Question Answering in automatic transcriptions of lectures (CHIL corpus)
- T2a: Question Answering in manual transcriptions of meetings (AMI corpus)
- T2b: Question Answering in automatic transcriptions of meetings (AMI corpus)
- T3a: Question Answering in manual transcriptions of broadcast news for French (ESTER corpus)
- T3b: Question Answering in automatic transcriptions of broadcast news for French (ESTER corpus)
- T4a: Question Answering in manual transcriptions of European Parliament Plenary sessions in English (EPPS English corpus)
- T4b: Question Answering in automatic transcriptions of European Parliament Plenary sessions in English (EPPS English corpus)
- T5a: Question Answering in manual transcriptions of European Parliament Plenary sessions in Spanish (EPPS Spanish corpus)
- T5b: Question Answering in automatic transcriptions of European Parliament Plenary in Spanish (EPPS Spanish corpus)

French broadcast news: the **ESTER corpus** (Galliano et al., 2006) is made of 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 Word Error Rates (WER = 11.0%, 23.9% and 35.4%). The manual transcriptions were produced by ELDA.

Spanish parliament scenario: the **TC-STAR05 EPPS Spanish corpus** (TC-Star, 2004 2008) is made of three hours of recordings from the European Parliament in Spanish. The data was firstly used in the TC-STAR project. There are 3 different automatic speech recognition outputs with different Word Error Rates (WER = 11.5%, 12.7% and 13.7%). The manual transcriptions were done by ELDA.

English parliament scenario: the **TC-STAR05 EPPS English corpus** (TC-Star, 2004 2008) is made of 3 hours of recordings from the European Parliament in English. The data was firstly used in the TC-STAR project. There are 3

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

different automatic speech recognition outputs with different Word Error Rates (WER = 10.6%, 14% and 24.1%) . The manual transcriptions were done by ELDA.

In the 2008 QAst evaluation, two kind of questions are considered : *factual questions* and*definition questions*. The factual questions are the same kind as the ones of the 2007 evaluation. In these questions, the answer to the search is a Named Entity (cf. section 3.1.). The definition question 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.*

In the 2008 evaluation, as in the 2007 pilot evaluation, an answer is basically structured as an [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 consists 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 transcription (the time segment *1019.228 1019.858*).

5. Conclusions

This paper has reported on Question Answering on Speech Transcriptions as defined in the pilot QAst evaluation track held in CLEF 2007, and described some of the plans for the future. The future evaluations are extending the QAst exercise to cover multiple languages (English, Spanish and French) and data types (European Parliament sessions, Broadcast News).A call for participation in the second QAst evaluation was recently announced as part of CLEF 2008 Multiple Language Question Answering (QA@CLEF) track.

6. Acknowledgments

This work was partly supported by the European Commission under the FP6 Integrated Project IP-506909 CHIL, the the OSEO under the Quaero program, and by the Spanish Ministry of Education and Science under the project TextMess. The authors are greatful to the FP6 Integrated Project IP-506811 AMI project for kindly sharing their data.

7. References

AMI. 2005. The AMI meeting corpus. http://www.amiproject.org.

Christelle Ayache, Brigitte Grau, and Anne Vilnat. 2006. Evaluation of question-answering systems : The French EQueR-EVALDA Evaluation Campaign. In *Proceedings of LREC'06*, Genoa - Italy, 24-26 May.

CHIL. 2004-2007. http://chil.server.de.

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

D. Giampiccolo, P. Forner, A. Peas, C. Ayache, D. Cristea, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, and

R. Sutcliffe. 2007. Overview of the CLEF 2007 Multilingual Question Answering Track. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, September.

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

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

N. Kando. 2006. Overview of the Sixth NTCIR Workshop. In *Proceedings of the 6th NTCIR Workshop Meeting*, Tokyo, Japan.

L. Lamel, G. Adda, E. Bilinski, and J.-L. Gauvain. 2005. Transcribing Lectures and Seminars. In *in InterSpeech'05*, Lisbon, Portugal.

D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. 2007. The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms. *Journal on Language Resources and Evaluation*, 41(3-4):389–407, December.

G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. 2007. CLEF 2007: Ad Hoc Track Overview. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, September.

QASTLE. 2007. http://www.elda.org/qastle/.

TC-Star. 2004-2008. http://www.tc-star.org.

J. Turmo, P. Comas, C. Ayache, D. Mostefa, S. Rosset, and L. Lamel. 2007. Overview of the QAST 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, September.

E. M. Voorhees and L. P. Buckland. 2007. The Sixteenth Text REtrieval Conference Proceedings (TREC 2007). In Voorhees and Buckland, editor, *NIST Special Publication 500-274*.