6

Spoken Question Answering

Sophie Rosset[1], Olivier Galibert[1,2], and Lori Lamel[1]

[1] *CNRS-LIMSI, Paris, France*
[2] *LNE, ?*

This chapter covers Question-Answering (QA) from spoken documents (referred to as QAst), but also beyond, where questions are also spoken. After a general introduction, Section 6.2 presents some specific aspects that must be considered when handling speech data for question answering. Then, Section 6.3 presents the main evaluation campaigns that have been carried out in the question-answering domain, the majority of which have addressed only written language. To the best of our knowledge, to date there has only been one evaluation for spoken QA. However, since the general problematics are the same, we believe it important to also present the larger view including written QA. The important considerations of QA evaluations are presented, followed by a detailed presentation of the QAst campaigns. Section 6.4 describes and compares different approaches and systems for QA, with a focus on approaches used for spoken language. This is followed by a review of recent projects or work addressing spoken QA. The chapter concludes with a discussion and some perspectives.

6.1 Introduction

Question-Answering systems can be seen as an extension of the Information Retrieval (IR) engines 15 which allow a user to search for information using a set of keywords. The result of the search is a set of documents, or links to documents, which the user needs to peruse to find the precise information wanted. In contrast, the question answering (QA) task consists of providing short, relevant answers to natural language questions which can be textual or spoken. Optionally the task can require the system to also return a document or document snippet supporting or even justifying the answer. Figure 6.1 illustrates the difference between IR and QA, which can be summarize by the following two points: First, the input is a natural-language question rather a keyword query; and second, the answer provides the desired information content and not simply a potentially large set of documents or URLs that the user must plow through. Progress in the QA domain can be observed via evaluation campaigns held since 1999 ((Dang et al. 2007; Forner et al. 2008; Mitamura et al. 2008; Voorhees and Tice 1999)).

Spoken question-answering is a new challenge for question-answering systems which generally deal with (well-formed) textual data and well-formed written questions. Spoken question-answering implies doing the search in spoken data and/or from spoken questions. This is a departure from much of the QA research carried out by natural-language groups, who have typically developed techniques for written texts that are assumed to have a correct syntactic and semantic structure. The structure of spoken language is different from that of written language, and some of the anchor points used in processing such as

Spoken Language Understanding Edited by Gokhan Tur and Renato De Mori

Information Retrieval
User query: building Eiffel Tower
System answers:
Document link: Building of the Eiffel Tower
Document snippet: Every metallic part of the Eiffel Tower is riveted...

Document link: Eiffel Tower - Wikipedia
Document snippet: Go to the *Tower building* (link to another page)
...

Question Answering
User query: When was the Eiffel Tower built?
System answer: 1887 - 1889
Document snippet: Built between 1887 and 1889 and followed by its inauguration at the universal exposition of 1889 in Paris, the Eiffel Tower nowadays symbolizes...

Figure 6.1 Looking to know the period when the Eiffel Tower was constructed.

punctuation must be inferred and are therefore error-prone. It is also necessary to deal with spoken-language phenomena including disfluencies, repetitions, restarts, and corrections. If automatic processing is used to create the speech transcripts, an additional challenge is dealing with the recognition errors. When dealing with speech data the response can be a short string, as is the case for text-based QA, or a brief audio segment containing the response.

A specific track – Question-Answering on Speech Transcripts (QAst) – has been proposed for now three years (2007, 2008 and 2009) in the Cross Language Evaluation Forum (CLEF) campaigns (Turmo et al. 2007a, 2009, 2008).

Question-answering in *spoken data* collections means that the answer has to be found in the audio data. These data can be of various types: broadcast news, meetings, seminars etc. The difficulty is quite dependent on the kind of data: broadcast news data are principally comprised of prepared speech and are quite similar to news texts; meetings and seminars contain spontaneous, interactive speech with a number of oral phenomena like speech repairs, hesitations etc, typical of this speaking style. Moreover, in real use-cases, spoken data are only accessible after having been processed by automatic speech recognizer (ASR) which means that the input to the QA engine is almost always an imperfect transcription of what was said. It is also well-known that ASR performance is dependent on the task. These factors increase the difficulty of the QA task.

Question-answering is not only a matter of the kind of data in which the answer has to be found, but also of the way the question is formulated. The input to spoken question-answering can be a written or a *spoken question* (which may or may not be interactive). Spoken questions are not necessary well-formed, at least not in the sense of written questions, with the syntax of oral language being quite different, in particular for questions, than that of written language. Of course, if the input is spoken, then there is also the need for automatic speech recognition, and the possibility that errors are introduced in the question (either by the user or the system).

A important question when working on (spoken) question-answering concerns the need of an understanding process. Why is understanding needed for question-answering? First of all, to find a precise answer to a precise question requires understanding the question. This means that the system

has to understand what kind of answer is expected, which in turn means understanding the important elements of the question. Moreover, it is also necessary to understand the documents in which the answer has to be found (answer extraction). An important consideration is that the same information can be formulated in many different ways, which must be taken into account when searching for an answer. In conclusion, question-answering requires working with a semantic level representation of both documents and questions.

6.2 Specific aspects of handling speech in QA systems

In this section, we present and illustrate the specific aspects of handling speech in QA systems, which entails working on two levels. The first level is developing or adapting methods, algorithms and tools to be efficient on transcribed speech. The second level is enhancing these approaches to handle, or at least be robust to, errors produced by automatic speech recognition systems.

Multiple aspects differ when working with spoken rather than with written language. First transcribed speech is structurally different than written texts. At the surface level one can notice the lack of punctuation and of clearly delimited frontiers between chunks, in particular numbers which are typically written out as words. For instance in *by trial two thousand five hundred twenty tests had failed* is the topic 20 tests, 120 tests, 520 tests, 2520 tests, or even a date followed by a number of tests? A human transcriber could add a comma to disambiguate this text or reconstruct the numbers in digits, but even humans can make an incorrect interpretation. Prosody may help in disambiguating such sentences, but current automatic speech recognition systems do not provide information at that level. Studies, for example (Liu et al. 2006), have addressed the use of prosodic information for the annotation (and handling) of disfluencies and sentence boundaries. Although widely acknowledged to be useful for a human reader (Jones et al. 2003) and often required for downstream processing (Lee et al. 2006), one of the reasons that ASR systems do not usually produce a punctuated output is that attempts to evaluate automatic punctuation have been inconclusive (Gauvain 2006). Another example illustrating the importance of punctuation is *the release of Christian Chesnot and Georges Malbrunot the prime minister nevertheless condemned ....* In this example, without the structural clues that would usually be present in a written text, it is hard to decide without world knowledge whether *Georges Malbrunot* should be attached with *Christian Chesnot* to *release*, or whether he should be associated with *prime minister*.

Lack of punctuation can also be problematic when extracting passages, since a large number of QA systems rely on the concept of a sentence and define passages as individual sentences. Without specific tokens delimiting sentences, such approaches are no longer viable. Another aspect is the deep structure of the word sequence. The syntax of speech is different than for written text, even if there are many commonalities. In particular the structuration is more local and dependencies generally link chunks which are closer in speech than in text (Miller and Weinert 1998).

Finally, the last aspect, and one not to be neglected, is that handling speech usually means handling the output of automatic speech recognition systems, along with their errors. When using the word error rate (WER) [1] to evaluate an ASR system, a mistake on a proper noun is considered as important as an incorrect pronoun-reference agreement. However, when using the ASR output as the input to another task such as QA, some errors are critical and others not important.

Figure 6.2 presents some example problems that can be encountered in ASR system outputs. We first notice that the system *ASR_C* does not provide true casing, which may complicate the task of Named Entity recognition. Comparing all ASR transcripts to the reference, it can be seen that the named entity *Kosovo* is not recognized by any of the systems, which makes it difficult – if not impossible – to answer the question *Which country are the prisoners from?*. The system *ASR_B* did not recognize the word *captives*,

[1] It has been reported that the overall word error rate is closely correlated with the error rate of named entities (Kubala et al. 1998), and the WER has remained the metric used to assess ASR system performance.

manual transcripts: one of the captives is from the Philippines one is from Kosovo and one Annetta Flanigan is from my constituency of Northern Ireland .

ASR A: one of the captives from the Philippines from possible on one another fun against from my constituency of Northern Ireland **ASR B:** one of the capitals from the for the previous ones from possible and one another for elegance from my constituency of Northern Ireland **ASR C:** ONE OF THE CAPTIVES FROM_THE FROM_THE PLAINS WARMS FROM POSSIBLE ON ONE ANOTHER FUN AGAIN THIS FROM MY CONSTITUENCY OF NORTHERN IRELAND

Figure 6.2 Examples of specificities of spoken language and automatic speech recognition outputs.

selecting the word *capitals* instead. Without that key element which matches the word *prisoners* part of the question, it is unlikely to find an answer in this passage.

As will be seen further in Section 6.4 these problems have been studied and taken into account by the participants to the QAst evaluation campaigns (Turmo et al. 2007a, 2009, 2008).

6.3 QA evaluation campaigns

This section gives an overview of the main Question-Answering evaluation campaigns, most of which were part of the TREC or CLEF benchmarks, and for the most part have focused on written language.

6.3.1 General presentation

Evaluations in the field of Question-Answering started with the Text REtrieval Conference (TREC) benchmarks, organized by the National Institute of Standards and Technology (NIST). The first QA task was introduced in 1999 ((Voorhees and Tice 1999)) and were subsequently organized annually until 2008 when TREC became TAC (Text Analysis Conference) with different objectives. In Europe, the Cross-Language Evaluation Forum (CLEF) was created in 2000 and introduced QA as one of the tasks in 2003 ((Magnini et al. 2003)). On the Asian front, NII Test Collection for IR systems (NTCIR) was created in 1999 with QA making a first appearance there in 2002 (Fukumoto et al. 2002). These three evaluation frameworks can be considered the most influential in the field. Of these evaluation campaigns, only one addressed the issue of QA and speech. QAst, one of the QA-related tracks of CLEF, proposed an evaluation with spoken documents and written questions from 2007 to 2009, adding spoken questions in 2009.

Table 6.3 shows some of the characteristics which have an impact on what needs to be done to answer questions. The top block of this table focuses on question type; the second block concerns the data collection and the last one concerns miscellaneous aspects.

Some of these characteristics are common between QAst and other QA evaluations, and are the focus of the following discussion. The characteristics that are specific to the QAst evaluation are fully described in Section 6.3.1.

Question types

When working on question-answering, the first question to answer is *what kind of questions should the system be able to answer?* For different question types, different algorithms are often used.

The simplest and most frequent question type is a factual question. These questions are questions for which the answer can be a single word or a multi-word expression, often is a named entity. An example

	TREC									QA@Clef Main Track							QAst			NTCIR				
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2003	2004	2005	2006	2007	2008	2009	2007	2008	2009	2002	2004	2005	2007	2008
Factual	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Simple definition		●	●								●	●	●	●	●							●		
Definitions					●													●					●	●
Why																							●	
How											●											●		
Yes/no																								
Open lists					●	●	●	●	●				●	●	●					●		●	●	●
Closed lists			◇	◇									●	●	●					●		●	●	
Follow-up			◇				●	●	●					●	●					●	◇	●		
Topics						●	●	●	●															
Information						●	●	●	●															
Spoken questions																			●					
Newspapers	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●					●	●	●	●	●
Speech																	●	●	●					
Law																●								
Wikipedia														●	●									
Blogs									●															
Question class given					●	●	●	●	●											●	●			
Multiple answers	●	●	●							●			●				●	●	●	●	●	●	●	●
Long answers	●	●	●							●	●													
Justification												●	●	●	●									
Translingual										●	●	●	●	●	●	●					◇	◇		●
Parallel docs.																●								
Temporal restriction												●	●	●	●									
Timecodes																		●	●					

◇: external task Sources:

- TREC: (Dang et al. 2006, 2007; Voorhees 2000, 2002, 2003, 2004; Voorhees and Dang 2005; Voorhees and Tice 1999, 2001).
- QA@Clef Main Track: (Forner et al. 2008; Giampiccolo et al. 2007; Magnini et al. 2006, 2003, 2004; Peas et al. 2008; Vallin et al. 2005).
- QA@Clef QAst: (Turmo et al. 2007a, 2009, 2008).
- NTCIR: (Fukumoto et al. 2002, 2004, 2007; Kato et al. 2004, 2005; Mitamura et al. 2008; Sasaki et al. 2005, 2007).

Figure 6.3 Table summarizing the main characteristics of the principal QA evaluations.

question is: *Who is the French president?* These kinds of questions are present in all QA evaluation campaigns.

There have been attempts to move from factual questions to so-called 'list' questions. These are factual questions for which the answer is a list of elements. In this category, two kinds of questions can be distinguished. Closed-list questions give the number of expected elements (*which are the three press companies that were subject to judgment by the Supreme Court?*. Open-list questions give no information about how many elements are expected (*what are the ingredients of Creme Anglaise?*. List questions have never been part of the QAst evaluation campaigns.

The second type of questions are 'Definition' ones for which two categories can be distinguished. For a general definition question, any kind of answer (named entity or not, simple word, phrase, complete sentence etc.) is possible. The second category is one in which only simplified definition questions have been proposed, as in (Vallin et al. 2005). These include questions of the type *who is Barak Obama?* and the answer can be a simple word, a multi-word expression or a named entity. This last category was present in TREC 2000 and 2001, in QA@CLEF 2004-2008 and in all QAst evaluation campaigns.

More complex question types also exist, such as *how*, *why* and *yes/no* questions. These questions are considered *complex* because the answer is not a word or a simple expression. For example, an answer to the question *How do dolphins get caught in driftnets?* could be *though intended to catch fish, the nets indiscriminately catch virtually all aquatic life including fish, whales, dolphins, sea turtles, and sea birds*. The *why* question *Why do college students eat poorly?* could answered by the following excerpt: *Stress, irregular schedules, parties, and the freedom to eat French fries smothered in cheese sauce*. In a similar vein, while yes-no questions can be answered with a simple word (yes or no), this is usually not considered satisfactory. For instance, a better answer to the question *Was Rosa Parks an African American?* would be a wikipedia citation such as *Rosa Louise McCauley Parks was an African American civil rights activist* ... rather than a simple *yes*.

In order to mimic interactivity, follow-up questions have been experimented with some evaluations (cf. (Giampiccolo et al. 2007; Kato et al. 2005; Voorhees and Dang 2005)). The idea was to measure how well systems could handle anaphora resolution. An alternative, proposed by (Voorhees 2004), introduced a set of questions linked to a topic. None of these complex questions were used in the QAst evaluation campaigns.

The 2009 QAst evaluation proposed natural, spontaneously spoken questions. It is the only evaluation which proposed such question types in an open domain. A more detailed discussion of these questions is presented in the Section 6.3.1.

Document types

Another important aspect of the question-answering evaluation is the collection of data in which the search has to be carried out. Of the different campaigns, only the QAst evaluation explicitly worked with spoken data. Most of the QA evaluation campaigns use data collections of newspaper documents. The corresponding spoken data are broadcast news documents which were used in some QAst evaluations. These documents have various advantages in the QA sense: first, they are relatively well formatted and clearly written (for broadcast news these contain primarily prepared speech); there is a lot of factual information and a fair amount of redundancy. Some evaluation campaigns included Wikipedia data in their data collections. The quantity of information is much bigger in Wikipedia than in typical two-year collections of newspaper documents, but the redundancy is a lot less. Other specific domain data collections have occasionally been used, for example, the JRC-Acquis in the QA@CLEF 2009 (see (Peas et al. 2008)).

The QAst evaluations have been organized with the objective of specifically working with spoken documents as further described next.

Spoken documents and spoken questions

Two different aspects have been investigated in the QAst evaluation campaigns: question-answering in spoken data and question-answering from spoken questions (in spoken data). The data collections used in the different editions of this evaluation campaign are described in Section 6.3.2, along with the different kinds of questions that have been proposed.

Metrics

There are several metrics typically used to evaluate question-answering systems. The simplest one is the *accuracy*, that is the ratio between the number of correct answers and the number of questions. When systems are allowed to return multiple answers, only the first one is taken into account when determining the accuracy. Letting CA_i be the rank of the first correct answer for the question i, defaulting to $+\infty$ if no answer was found:

$$\text{accuracy} = \frac{\#\text{CA}_i = 1}{\#\text{questions}} \tag{6.1}$$

Stopping at the first answer is somewhat limiting. A direct extension is the *top-n* accuracy, which takes into account correct answers between ranks 1 and n:

$$\text{top-}n = \frac{\#\text{CA}_i \leq n}{\#\text{questions}} \tag{6.2}$$

In practice, in addition to $n = 1$ which is also referred to as the 'raw' accuracy, *top-n* tends to be used with n equal to the maximum number of answers the systems are allowed to return. This use turns the *top-n* measure into a kind of *recall*.

The system is required to put the most probable answers at the top of the list. To measure the quality of that classification, the *Mean Reciprocal Rank* (MRR) is often used. The first correct answer gives a score for the question equal to the inverse of its rank. The lack of a correct answer is equivalent to an infinite rank and hence a score equal to zero. The final score is the arithmetic mean of the individual question scores:

$$\text{MRR} = \frac{\sum \frac{1}{\text{CA}_i}}{\#\text{questions}} \tag{6.3}$$

Accuracy, MRR and Top-n taken together are a triplet of values which give an idea of the quality of a system and its evolution potential. They are also useful to compare results between multiple versions of the same system. These three metrics were used in the QAst campaigns. Other metrics exist to handle other question types such as lists or extra information such as confidence levels, but they were not pertinent for the QAst campaigns.

6.3.2 Question Answering on speech transcripts: evaluation campaigns

The Question-Answering on Speech Transcripts evaluation campaign (QAst) was created in 2007 to investigate the problem of question-answering on speech data ((Turmo et al. 2007b)).

The data for the QAst evaluation campaigns was derived from five different resources listed below, covering spontaneous speech, semi-spontaneous speech and prepared speech. Data from the first two corpora (CHIL, AMI) were used in the 2007 and 2008 editions ((Turmo et al. 2007a, 2008)). Data from the other corpora were used in the 2008 and 2009 editions ((Turmo et al. 2009)).

- The **CHIL corpus**[2]: The corpus contains about 25 hours of speech, mostly spoken by non-native speakers of English, with an estimated ASR WER of 20%.
- The **AMI corpus**[3]: 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** ((Galliano et al. 2006)) 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 reference transcriptions were produced by the Evaluations and Language resources Distribution Agency (ELDA).
- Spanish parliament: The **TC-STAR05 EPPS Spanish corpus** ((TC-Star 2004-2008)) 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 reference transcriptions were produced by ELDA.
- English parliament: The **TC-STAR05 EPPS English corpus** ((TC-Star 2004-2008)) contains 3 hours of recordings from the European Parliament in English. The data was used to evaluate 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 reference transcriptions were produced by ELDA.

The spoken data cover a broad range of types, both in terms of content and in speaking style. The Broadcast News and European Parliament data are less spontaneous and less interactive than the lecture and meeting speech, typically being prepared in advance and as such 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 prepared speech 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 had been developed for these languages and tasks in the context of European or national projects ((Galliano et al. 2006; TC-Star 2004-2008)).

Questions and answer types

Two kinds of questions were considered: *factual questions* and *definition questions*. To the first question type, the expected answer of the search is a Named Entity. The definition questions are questions such as *What is the Vlaams Blok?* and the answer can be anything. In this example, given the data collection (see Figure 6.4), the answer could be *a criminal organization*. The definition questions can be further subdivided into the following types:

- **Person:** question about someone
  Q: *Who is George Bush?*
  R: *The President of the United States of America.*
- **Organization:** question about an organization
  Q: *What is Cortes?*
  R: *Parliament of Spain.*

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

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 organization*

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** organization 0.5847
20041115_1705_1735_EN_SAT 1 1020.938 0.100 and 0.9747
(...)
Answer: 1019.228 1020.698

Figure 6.4 Example query *What is the Vlaams Blok?* and response from manual (top) and automatic (bottom) transcripts. The CTM file format is *document id*, *channel number*, *temporal position*, *duration*, *word* and *confidence score*

- **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*.

An answer was given as a (answer string, document id) pair, where the answer string contains nothing more than the full and exact answer, and the document id is a unique identifier of the document supporting the answer.

For the tasks using automatic speech transcripts, the answer string specified the <start-time> and the <end-time> of the answer in the signal. Figure 6.4 illustrates this point comparing the expected answer to the question *What is the Vlaams Blok?* in a manual transcript (the text *criminal organization*) and in an automatic transcript (the time segment *1019.228 1020.698*).

6.4 Question answering systems

6.4.1 General overview

Question-Answering systems are generally organized as shown in Figure 6.5. They usually start by preprocessing documents prior to indexing them. This preprocessing stage can be separated into two parts. The first part, which does not always exist, is a form of language analysis which attempts to extract some structured information from the documents. This analysis can be relatively simple, for example determining only the levels of Parts-of-Speech and Named Entities, such as in (Comas et al. 2007; Molla et al. 2007, 2006). These relatively simple levels of analysis are particularly relevant for the Question-Answering on Speech Transcripts task, as reported in (Turmo et al. 2007a, 2008), probably because of

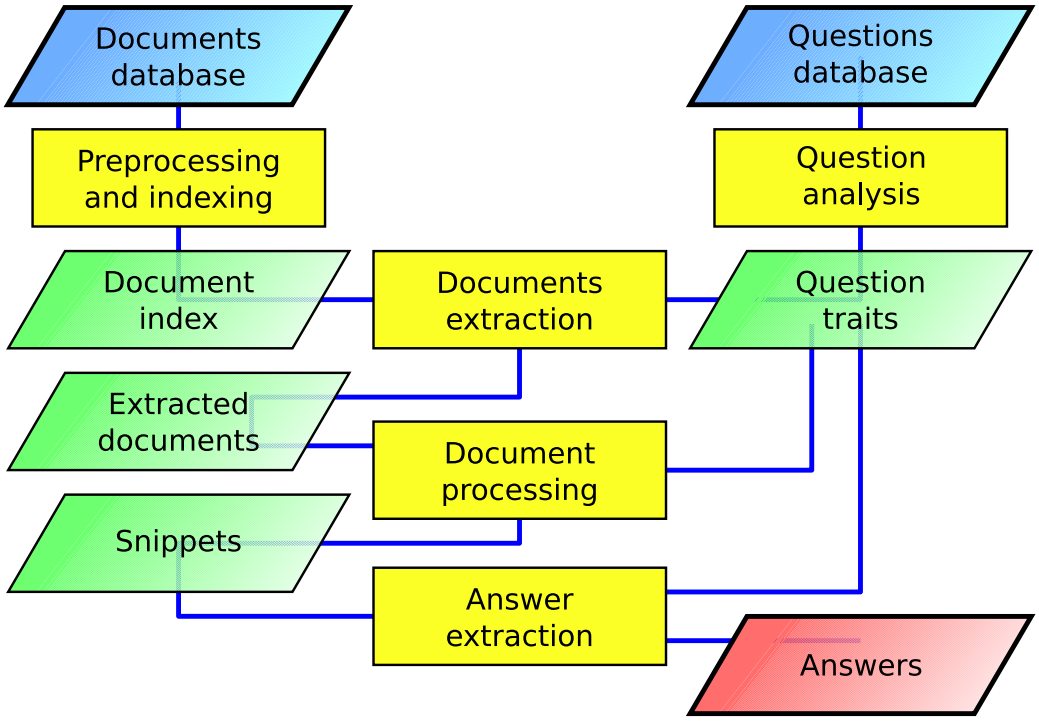

Figure 6.5 Overview of typical QA system architecture (inspired by (Ligozat 2006))

the difficulty of adapting more complex analyses, such as syntactic analysis, to speech. In the case of processing clean texts, the analysis may try to reach deeper levels through syntactic and even semantic analysis (Hickl et al. 2006; Laurent et al. 2006; Neumann and Wang 2007).

The second part of the document preprocessing concerns reformatting prior to indexation. Systems designers often do not want to work with whole, raw documents as the indexation unit, so it is common to split the documents into smaller units. Sentences can be a preferred unit, or small blocks of them (Laurent et al. 2006). In the speech case, sentence-like blocks can be built (Krsten et al. 2008).

Indexation for subsequent retrieval is then done by a search engine. A very popular choice is LUCENE (Apache 2007), an open-source search engine currently developed by the Apache foundation (Comas et al. 2007; Neumann and Wang 2007). Managing Gigabytes (MG) is also used, for instance by (Grau et al. 2005), but since development has stopped its popularity has been declining. A number of systems use their own indexation and retrieval engines (Rosset et al. 2008), in particular those who do linguistic analysis in the preprocessing and want to be able to search in the results; (Laurent et al. 2006) with its deep analysis and associated specialized indexing is a good example.

In the specific case of speech, where the speech recognition engine may generates a transcription with errors, two systems (Turmo et al. 2009, 2008) tried to use a phonemic based search. The underlying assumption is that the speech recognition engine errors can derail the question answering process if they occur on critical keywords (proper names for instance) but that the phonetic transcription of the erroneous words is not far from the true content of the audio signal. This approach was promoted by (Clements et al. 2001) to locate information in audio archives. For Q&A (Comas and Turmo 2008a) used a specific

retrieval engine, PHAST (PHonetic Alignment Search Tool), to retrieve near sequences of phones built for the question keywords in the documents using a similarity measure. This approach called *Keyword Spotting* in the speech retrieval domain.

Once retrieval is ready to run, the work turns towards the questions. The question analysis aims at handling two problems: the first is to detect which information in the question has to be found in the documents. This information often takes the form of keywords and named entities, but sometimes is a syntactic or semantic relation. This first analysis is usually very close to what is done for document analysis, in order to be able to compare the information found in questions and documents, a necessity for the following extraction steps. The second problem is to predict what type of answer to the question is expected (Pardio et al. 2008). The answer type is usually a named entity category (person, location...), but can be more specific or have a broader coverage when more advanced taxonomies are used. This process is often called *question classification* or *expected answer type detection*. Question classification is an important task, allowing unrelated documents to be filtered out or applying specific handling for snippet or answer extraction. Question classification allow predicting what kind of precise answer is to be searched for, and what constraints the question imposes on possible answers. Most systems are based on pre-defined question categories (see for example (Amaral et al. 2005; Hacioglu and Ward 2003; Li and Roth 2002)).

Question categorization can be done using simple patterns (Monz and de Rijke 2001) or using machine learning approaches (Ferres et al. 2004). For example, (Hacioglu and Ward 2003) used Support Vector Machine (SVM) classifiers to learn models able to classify questions and predict the expected answer type. (Li and Roth 2002) used a classifier based on the Sparse Network of Winnows (SNoW) learning architecture. Using a sparse network of linear units, a hierarchical classifier was constructed. The hierarchy contained 6 coarse classes and 50 fine-grained classes.

The results of the information extraction part of the question analysis are given to the search engine which retrieves whole documents or snippets, as defined by the indexation. A complementary analysis can be done on the results (i.e. the returned documents), similar to the preprocessing described previously. Which part of the general document analysis should be done before or after indexing is essentially an engineering question, balancing preprocessing time, response time, indexing complexities and scaling issues. In any case, the final step is to extract candidate answers from these analyzed snippets and rank them. In most of the cases, candidates answers are chosen as word or phrases annotated with the expected type for the answers. A score is given to each of them which can be based on answer-keyword distance (Pardio et al. 2008), density (Comas and Turmo 2008a; Gillard et al. 2006), or even syntactic similarities and dependency relations (Bouma et al. 2005). The system then ranks the answers according to the scores.

This very general structure covers most of the linguistically motivated approaches to question answering. Some alternative methodologies exist, which rely on a minimal or even no linguistic knowledge. (Berger et al. 2000) for instance uses statistical models based on co-occurrence measurements. (Ittycheriah and Roukos 2002) added two enhancements to this approach. First, a snippet selection method based on IBM Model1 translation models, considering pertinent justification snippets as ”translations” of the question and using the estimated translation probability as a snippet sentence pertinence score. Second, a set of automatically extracted answer patterns is proposed to improve over the co-occurrence measurements. (Whittaker et al. 2007) applied similar approaches in the QAst 2007 evaluation.

The evaluation campaigns show that the best performing systems are usually based on a deep linguistic analysis. The LCC system (Moldovan et al. 2002) is a good example of such a system. In the 2005 TREC evaluation, this system obtained the best results with 71% correct answers on factual questions. In comparison, the fully stochastic Tokyo Institute system (Whittaker et al. 2005), obtained only 21% correct answers in the same evaluation (Voorhees and Dang 2005).

6.4.2 *Approaches used in the QAst campaigns*

Most of the systems participating in the QAst evaluations used a standard architecture. The main differences concern the way speech data (transcriptions) was dealt with. There are several aspects where question answering can be adapted to spoken language.

The difficulties of transcribed speech can be handled at two different steps in the system: when analyzing the documents or when matching documents and questions. In other words, specific treatment can be carried out at indexation time or at information retrieval time. In the first case, the aim is to take into account the structural specificities of speech. In the second, the issue is more about handling and compensating for speech recognition errors.

In the following sections we describe the adaptation for handling speech that has been done at the different stages of question-answering systems in the QAst evaluations: question and document processing, and information retrieval. The answer extraction and ranking did not require specific adaptation. Three participants proposed adaptations to deal with speech: the Universitat Politecnica de Catalunya (UPC), the Instituto Nacional de Astrofisica, Optica y Electronica (INAOE) and the Laboratoire d'Informatique pour la Mcanique et les Sciences de l'Ingnieur (LIMSI).

Question and document processing

Working on spoken data does not fundamentally change the way systems work, although some adaptations are required. First, not all text analysis approaches are robust on transcriptions of speech, and in particular on automatic speech recognition outputs. Second, alternative approaches can try to mitigate the impact of speech recognition errors.

Document processing can be adapted in different ways, from enrichment of the document representation to an adapted analysis. Analysis can concern both document and question processing, and usually takes the form of simple POS tagging and Named Entity detection. Most QAst systems only apply such an analysis to the documents and not to the questions. The LIMSI systems (Bernard et al. 2009; Rosset et al. 2007, 2008) and (Neumann and Wang 2007) system are exceptions. All other systems try to classify the question in order to extract the answer type and to extract the keywords used for information retrieval (see for example (Comas et al. 2007; Krsten et al. 2008)).

(Neumann and Wang 2007) carries out a complete question analysis leveraging from the fact that the QAst 2007 questions were provided in a written form. Their question parser computes for each question a syntactic dependency tree which also contains the recognized named entities. This parser is the same one that is used for text-based question-answering. Based on this analysis, their system produces a list of expected answer types.

It can be observed that Named Entity detection plays a central role in question-answering systems, and in particular in the QAst systems. The task limits the acceptable answers to named entities of a fixed type set, making detection of these entities in the documents particularly valuable. (Neumann and Wang 2007) explored the use of different existing Named Entity and POS taggers, POS tagging often being the first step for a NE tagger, and concluded that the models trained for written language are not very efficient on spoken language, and that linguistic annotated spoken data is needed.

To address this problem, the UPC team ((Comas and Turmo 2008a, 2009)) used development data to train a specialized model for English. This approach gave good results (F-measure of 75) on a set-aside test portion of the AMI meeting development data but not for the CHIL lecture data with a F-measure of 33. For the latter, the outputs of different systems (the one trained on the development data, a model trained on ConLL English corpus[4] and a rule-based system) were merged. This combination did not improve the precision but obtained better recall. The features used by the system and applied on manually transcribed documents are words, lemmas, POS tags, word affixes, flags indicating the presence of

[4] http://cnts.ua.ac.be/conll2002/ner

numerals and capitalization, and n-grams of these features. Gazetteers were also applied to provide an additional set features.

These features are relatively standard in the community. UPC also developed specific models to deal with ASR outputs, by expanding the features used by the classifiers to include with phonetic attributes. The basic arguement for this is that with ASR the phonemic structure tends to be maintained even if the word sequence is incorrect. For example, *Sun* can be misrecognized as *some* which while obviously an error is still phonemically similar. An unsupervised hierarchical clustering algorithm was used to group tokens based on the similarity of their phonetic features. The cluster of each token is added as a feature during training of the Named Entity recognition model. For the Spanish data, in 2008, a system was trained on the ConLL Spanish corpus. This system is not well suited to the QAst task because the corpus covers only three different types of named entities: *person*, *location* and *organization*). For the 2009 edition, the same strategy of system combination and specific models trained for ASR outputs was applied to develop a model for Spanish.

LIMSI ((Bernard et al. 2009; Rosset et al. 2007, 2008; Toney et al. 2008; van Schooten et al. 2007)) considers that the same analysis needs to be performed on questions and documents. Moreover, the team argues that because the speech transcriptions are the output of automatic speech recognizers they can benefit from some of the work already done when training ASR language models: words are clearly delimited, abbreviations are expanded which removes some of their inherent ambiguity, and uppercase, when present, is limited to proper nouns and acronyms. Ideally, the text at the entry of an analysis step would combine the advantages of both spoken and written text: words separated from the punctuation and from each other, uppercase only on proper nouns and acronyms, the presence of punctuation to allow splitting of sentences, etc. Therefore, a *normalized* form was defined which is the produced by the first stage of any of their systems, whether they apply to speech transcripts or to texts. This normalization stage includes the following processing steps:

1. Separating words and numbers from punctuation.
2. Reconstructing correct case for the words.
3. Adding punctuation.
4. Splitting document into sentences at period marks.

The first step relies on a series of algorithmic and regular-expression based transformations. Steps 2 and 3 are done simultaneously using a 4-gram language model trained on (reasonably) close written texts with punctuation and appropriate casing (on proper nouns and acronyms only) (Dchelotte et al. 2007). A word graph containing all possible punctuation and casing hypotheses is generated and the most probable one is selected via the language model. The result of the normalization is passed to the analyzer, which uses a multi-level approach to detect and type phrases according to a hierarchical taxonomy. For the French language, the analyzer detects approximatively 300 different types, spanning not only classic named entities and several extensions but also morphosyntaxic chunks and dialog acts. The analyzers for English and Spanish are partial adaptations of the French one.

To specifically handle the automatic speech recognition errors, INAOE (Reyes-Barragan et al. 2009), has proposed to enrich the request with a phonetic representation of the words in addition to traditional lemmas and other lexical derivatives. They used the Soundex (Odell and Russell 1918) code to derive phonetic forms. Since this representation having is sed for information retrieval, more details are provided in the the next section.

To summarize, the three approaches have been proposed to better handle spoken language. These can be used alone or in combination. The first one aims to adapt models designed for written text to speech transcripts. The second approach is to develop a common representation for documents independent of

whether their origin is speech or text, and to build an appropriate analysis. The third one is to enrich the representation with a phonetic layer.

Document processing can involve another operations, for instance, often the documents are split in shorter units. Sentences can be the preferred unit or small blocks of sentences. In the speech case, sentence-like blocks can be automatically identified. (Krsten et al. 2008) uses punctuation marks present in the manually transcribed documents. (Neumann and Wang 2007) used the sentence splitter of the OpenNLP tool, based on maximum entropy modeling, to identify sentence boundaries using a standard language model optimized for written documents.

Information retrieval

This step is one of the most important in a question-answering system. In the case of question-answering on speech transcripts, two original approaches have been proposed. The first one based on a phonetic codification was investigated by (Reyes-Barragan et al. 2009). The second one was explored by (Comas and Turmo 2008a, 2009; Comas et al. 2007) who proposed and developed a specialized Information Retrieval engine relying on phone similarity.

Phonetic codification
INAOE made use of the Soundex (Odell and Russell 1918) system. Designed only for the English language, the algorithm is based on a *it place of articulation* phonetic classification of human speech sounds (bilabial, labiodental, dental, alveolar, velar and glottal). Every word is transformed into a code composed of a initial letter, the first of the word, followed by a series of digits representing classes of sounds. A complete description of the algorithm is available in (Reyes-Barragan et al. 2009). The following example illustrates its use:

> **ASR output:** I am starting to work I do not know what <NOM> Frank Sinatra </NOM> must have felt like as his fellow appearances decided on in the seventies
> **After elimination of stopwords:** starting work <NOM> Frank Sinatra </NOM> felt like fellow appearances decided seventies
> **Phonetic codification:** S36352 W62000 F65200 S53600 F43000 L20000 F40000 A16522 D23300 S15320
> **Enriched representation:** starting work <NOM> Frank Sinatra </NOM> felt like fellow appearances decided seventies S36352 W62000 F65200 S53600 F43000 L20000 F40000 A16522 D23300 S15320

Phonetic similarity
Alternatively, UPC (Comas and Turmo 2008a, 2009; Comas et al. 2007) proposed to handle the speech recognition errors in the information retrieval model. The authors built the *PHAST* IR engine which has the capability to handling errors using phone similarity measures. PHAST uses pattern-matching to find short phone sequences (the keywords in the requests) in large phone sequences (the documents), and sorts them using a similarity measure. This is reminiscent of a traditional *word spotting* technique. The passage selection algorithm is then applied using the words found in that way by PHAST (Comas and Turmo 2008b).

6.4.3 QAst campaign results

This section highlights the results of the QAst evaluation campaigns held from 2007-2009. As described in Section 6.3.2, two different kinds of speech data have been used in QAst evaluation campaigns: prepared speech and spontaneous speech. Moreover, in the 2009 QAst evaluation campaign, spoken questions have been introduced.

	2007				2008			
	Acc.		MRR		Acc.		MRR	
	best	worst	best	worst	best	worst	best	worst
CHIL man.	51.0%	5.0%	0.53	0.09	41.0%	16.0%	0.45	0.16
CHIL ASR	36.0%	2.0%	0.37	0.05	31.0%	27.0%	0.34	0.30
AMI man.	25.0%	16.0%	0.21	0.06	33.0%	26.0%	0.40	0.29
AMI ASR	21.0%	6.0%	0.22	0.10	18.0%	14.0%	0.20	0.18

Figure 6.6 Best and worst accuracy and MRR on spontaneous speech data for 2007 and 2008 evaluation campaign

Speech style

The spontaneous speech data used in the QAst 2007 and 2008 campaigns were:

- The **CHIL**[5] **corpus** with about 25 hours of speech, mostly from non-native speakers of English, with an estimated ASR Word Error Rate (WER) of 20%.
- The **AMI** [6] **corpus** containing about 100 hours of speech, with an ASR WER of about 38%.

Table 6.6 gives an overview of the results obtained by the participating systems in 2007 and 2008 on spontaneous speech. For each condition, the performance of the best and worst system are shown.

Three data sets comprised of (primarily) prepared speech were used in QAst 2008 and 2009 evaluations:

- **French broadcast news data:** containing about 10 hours of broadcast news in French.
- **Spanish EPPS data** comprised of three hours of recordings from the European Parliament in Spanish.
- **English EPPS data** consisting of 3 hours of recordings from the European Parliament in English.

For each data set automatic transcriptions from 3 different automatic speech recognition systems with different error rates were used.

Table 6.7 gives an overview of the results obtained in the 2008 and 2009 evaluations on prepared speech, showing the best and worst results among the participating systems.

These two tables give an overview of the results obtained by multiple systems on a reasonably standard QA task which required searching in transcribed speech data. They also show the difference in performance when searching in error-free data (manual transcripts) and in data with ASR errors. Globally, the results are inferior to what the best systems obtain in clean textual data such as newswire and newspapers, for instance in the TREC (Voorhees 2002; Voorhees and Tice 2001) and QA@CLEF (Magnini et al. 2006; Vallin et al. 2005) evaluations. Although there is a clear loss of performance when switching from manual to automatic transcriptions, the speech type does not seem to be a significant factor in that loss. The results on the CHIL and AMI data are similar to those for the English European Parliament data.

Question style

One of the aims of the 2009 QAst evaluation was to compare using spoken and written questions. Table 6.8 shows the results obtained on spoken *vs* written question on the manually transcribed data collections.

[5] http://chil.server.de
[6] http://www.amiproject.org

	2008				2009			
	Acc.		MRR		Acc.		MRR	
	best	worst	best	worst	best	worst	best	worst
EPPS-E man	34.0%	20.0%	0.42	0.21	28.0%	5.0%	0.36	0.08
EPPS-E asr_1	30.0%	7.0%	0.33	0.10	26.0%	3.0%	0.31	0.07
EPPS-E asr_2	20.0%	10.0%	0.24	0.12	21.0%	3.0%	0.25	0.06
EPPS-E asr_3	19.0%	10.0%	0.23	0.12	25.0%	5.0%	0.28	0.11
EPPS-S man	31.0%	7.0%	0.35	0.09	36.0%	14.0%	0.45	0.20
EPPS-S asr_1	24.0%	3.0%	0.26	0.04	27.0%	6.0%	0.32	0.07
EPPS-S asr_2	19.0%	4.0%	0.22	0.05	25.0%	7.0%	0.29	0.10
EPPS-S asr_3	23.0%	2.0%	0.25	0.02	23.0%	9.0%	0.28	0.11
BN-fr man	45.0%	42.0%	0.49	0.47	28.0%	27.0%	0.39	0.38
BN-fr asr_1	41.0%	-	0.45	-	29.0%	-	0.37	-
BN-fr asr_2	25.0%	-	0.30	-	27.0%	-	0.32	-
BN-fr asr_3	21.0%	-	0.24	-	23.0%	-	0.28	-

Figure 6.7 Best and worst accuracy and MRR on prepared speech data for 2008 and 2009 QAst evaluation campaigns.

Data	Question types	Acc.		MRR	
		best	worst	best	worst
EPPS-E	written q.	28.0%	5.0%	0.36	0.08
EPPS-E	spoken q.	26.0%	3.0%	0.34	0.06
EPPS-S	written q.	36.0%	14.0%	0.45	0.20
EPPS-S	spoken q.	36.0%	17.0%	0.45	0.22
BN-fr	written q.	28.0%	27.0%	0.39	0.38
BN-fr	spoken q.	28.0%	28.0%	0.39	0.39

Figure 6.8 Best and worst accuracy and MRR given the question style during the 2009 evaluation campaign

It can be seen that there is almost no difference in the results with written or human transcripts of spoken questions. Looking more closely at all of individual system results and not only the best and worst ones (see (Turmo et al. 2009) for more information), five of the seven systems participating in the English task had a small loss of 2% to 5% when using transcripts of the spoken questions instead of written ones. The other two systems had a larger losses of 13% and 16%.

Discussion about results and approaches

Figures 6.9 to 6.11 show the results obtained by the INAOE, LIMSI, UPC systems in the QAst 2008 and 2009 evaluations. As previously described, these 3 systems are the ones which proposed specific approaches to handle question answering in speech data. Moreover, the best result was always obtained by one of the three in each task of these evaluations.

For the CHIL and AMI tasks (Figure 6.9 featuring spontaneous speech, the LIMSI system obtained the best results on manual transcriptions while UPC in the best on automatic transcriptions. For the EPPS and BN tasks which feature prepared speech, the INAOE English system achieved the highest precision level on the manual transcriptions with an identical *MRR* as the LIMSI system. When it comes to automatic

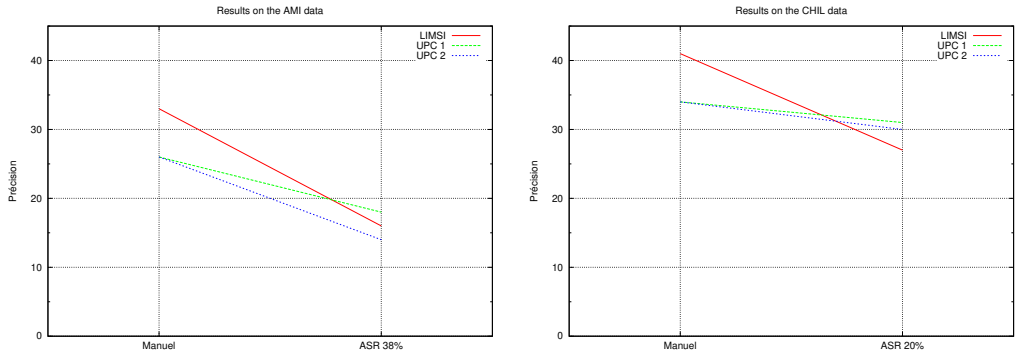

Figure 6.9 Results of the UPC and LIMSI systems on the CHIL and AMI tasks (2008).

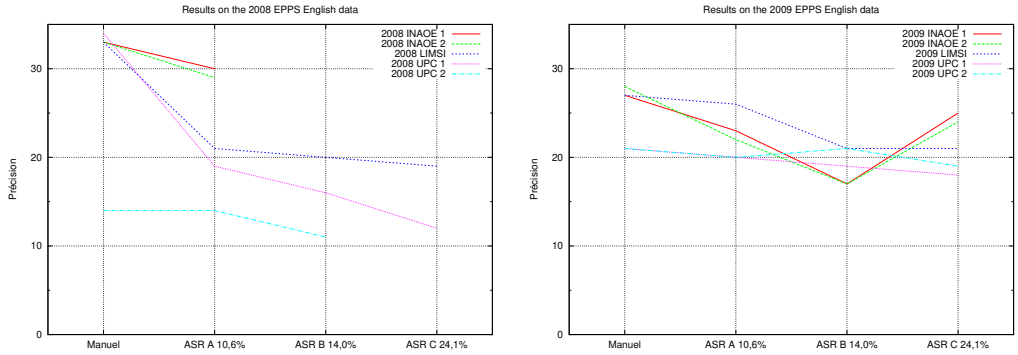

Figure 6.10 Results of the INAOE, LIMSI and UPC systems on the EPPS English task in 2008 and 2009.

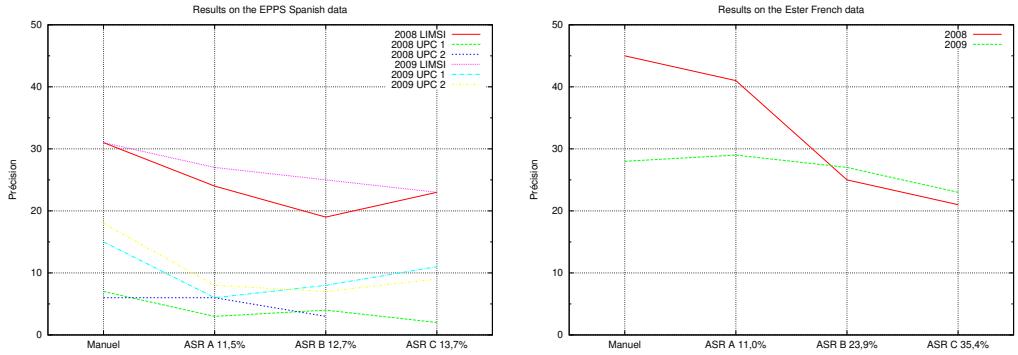

Figure 6.11 Results of the LIMSI and UPC systems on the EPPS Spanish and Ester French tasks in 2008 and 2009.

transcriptions, the LIMSI gets the best results for transcripts with the lowest two error rates (10.6% and 14.0%). INAOE proposed an alternative, *multi-ASR* approach for the third error rate, using information extracted with their Named Entity Recogniser from all of the ASR outputs, while still considering the third ASR output as the primary one. While that method is realistic in applicative setup when multiple ASR systems are available, it precludes measuring the impact of ASR errors. For Spanish LIMSI achieves the best results on both manual and automatic transcriptions.

LIMSI was the only participant for the French Ester BN task, so it is difficult to conclude much about the results. The performance on this task is globally among the best for all tasks, but it is difficult to ascertain if the is due to the approaches, to better processing and analysis for the French language, experience with the task or simply the task difficulty.

What can we learn from these results? First, we see that normalizing the documents of different origins to a unique surface representation seems to be beneficial. The system using that approach obtains close to or the best results on the manual transcriptions, and seems to be somewhat robust on the automatic transcriptions given that nothing specific is done. We can see that doing something specific at the analysis level, such as training a named entity extractor on ASR outputs as UPC did, gives an undeniable advantage on lectures and meetings where speech recognition errors are numerous.

Enhancing the requests or the documents with phonetic-level information does not seem to result in better performance. INAOE, which experimented with such an approach using the Soundex code to add phonetic class attributes to documents and requests, did not measure any improvement using it (Reyes-Barragan et al. 2009). Similarly, no matter the task or the language, the UPC system which did not use phonetic proximities for information retrieval ended up with better results than the one which did (Comas and Turmo 2008a, 2009). Nevertheless, it is intuitive that explicitly handling speech recognition errors one way or another should improve the robustness of the QA systems when the answer needs to be found in automatically transcribed documents. The main issue is determining how. Unfortunately most of the methods tried thus far, especially the phonetically-motivated ones, do not seem to enhance the results and even sometimes add noise, reducing performance.

6.5 Projects integrating spoken requests and question answering

Compared with text-based Q&A, there have been only a few studies searching in spoken documents and even fewer using spoken queries.

Most of the research projects aiming at integrating speech recognition and question answering systems made use of read questions. (Hori et al. 2003), for instance, developing on spoken and interactive Japanese question-answering system, used as input 69 questions read aloud by 7 male speakers. Similarly, (Gonzalez-Ferreras et al. 2008) asked speakers to read the questions of CLEF 2005. Such a setup reduces the intrinsic variability of speech, allowing the focus to be on the difficulties arising from automatic speech recognition (ASR) errors. And these difficulties can be significant: the results of the SPIQA system ((Hori et al. 2003)) dropped from an MRR of 0.43 to 0.25 when using ASR. (Gonzalez-Ferreras et al. 2008) reported a similar effect with their results dropping from 30.5% to 23.0% when using an ASR systems with a WER of 15.2%.

(Harabagiu et al. 2002) went further in their experiments. When they used read TREC-8 questions (30% WER for the ASR system) for their evaluations, the MRR dropped from 0.76 to 0.07. They then worked on the integration between Q&A and ASR and increased the MRR to 0.41 through a combination of filtering and language model selection, directed by knowledge dynamically extracted by the Q&A system. An additional gain of 0.07 was obtained if the system was allowed to ask the user for additional precisions and clarifications. This was the first step towards an interactive Q&A system.

(van Schooten et al. 2007) went then further by being first an interactive spoken system and second a Q&A service. The system was built around a dialog core with real-time ASR allowing spontaneous exchanges between a human and the computer. The system had to extract pertinent information to

build a request, information which was often spread across multiple utterances. The request was then communicated to an open domain Q&A system. One aim of the approach was to compensate for speech recognition errors via interaction by using an implicit confirmation request while running the Q&A system in parallel.

Although not strictly a QAst task, the distillation task, introduced in the GALE program (http://www.darpa.mil/ipto/programs/gale/gale_approach.asp) aims to use automated language technolgoies to extract relevant information in foreign language audio or text documents, and present it to users in English. Queries are specified in English and systems are to report relatant and non-redundant information, along with supporting documents (in English and in the source language). Different template types for queries were defined (in Phase 2 there were 17 different templates), and human references consist of nuggets of information extracted from snippets (Babko-Malaya 2008; Babko-Malaya et al. 2010). Performance is measured in terms of information content and document support.

Recently the Deep Q&A research project at IBM has received a lot of media attention, with the Watson system Jeopardy challenge, in which the computer will compete against some of the world's best Jeopardy! quiz show contestants. The game requires contestants to quickly respond to questions or clues, making fine distinctions of meaning and inferences about relationship between words and information content. While Watson can make use of massively parallel hardware and huge stores of information, it will not be connected to the Internet, thus needing to rely only on its memory and pre-coded logic. Although human contestants will respond to spoken questions, Watson will have typed ones, both to save time and avoid any potential ASR errors.

6.6 Conclusions

Spoken language question-answering is a newly emerging field, covering at the same time searching for answers in spoken documents and asking questions vocally. It builds upon over 20 years of research in speech recognition and spoken language processing, and a decade of in text-based Q&A. To date most research has addressed the first aspect, that is locating information in spoken documents, with less effort having been directed to using spoken queries. This latter aspect has been extensively addressed in spoken language understanding systems, typically for limited domains (travel or tourist information, for example). There are many potential applications of spoken Q&A technology for any tasks involving audio data mining or speech analytics. Speech recognition errors remain a problem, in particular if an error is made on an important word. Since the goal is to give a precise answer to a precise query, ASR errors may be more problematic for Q&A system than for IR ones, where a document may contain multiple instance of the information, one of which may be located. As speech recognition and search technologies improve, so will the potential for open-domain, interactive spoken Q&A systems. In additional to refining the search, natural interaction with a Q&A system can help improve speech recognition for voice queries by allowing the user to reformulate or clarify their questions.

References

Amaral C, Figueira H, Martins A, Mendes A, Mendes P and Pinto C 2005 Priberam's question answering system for portuguese *Working Notes for the CLEF 2005 Workshop*, Vienna. Austria.

Apache 2007 Apache lucene, an overview http://lucene.apache.org/java/docs/.

Babko-Malaya O 2008 Annotation of nuggets and relevance in gale distillation evaluation In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (ed. (ELRA) ELRA), Marrakech, Morocco.

Babko-Malaya O, Hunter D, Fournelle C and White J 2010 Evaluation of document citations in phase 2 gale distillation *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Berger A, Caruana R, Cohn D, Freitag D and Mittal V 2000 Bridging the lexical chasm: statistical approaches to answer-finding *Proceedings of the 23rd annual internation ACM SIGIR conference on research and develomment in information retrieval*, Athens, Greece.

Bernard G, Rosset S, Galibert O, Bilinski E and Adda G 2009 The limsi participation to the qast 2009 track *Working Notes of CLEF 2009 Workshop*, Corfu, Greece.

Bouma G, Mur J, van Noord G, van der Plas L and Tiedemann J 2005 Question answering for dutch using dependency relations *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria.

Clements M, Cardillo P and Miller M 2001 Phonetic searching vs large vocabulary continuous speech recognition: How to find what you really want in audio archives *Proceedings, Conference of Applied Voice Input/Output Society*, San Jose, CA.

Comas P and Turmo J 2008a Robust question answering for speech transcripts: Upc experience in qast 2008 *Working Notes of CLEF 2008 Workshop*, Aarhus, Denmark.

Comas P and Turmo J 2009 Robust question answering for speech transcripts: Upc experience in qast 2009 *Working Notes of CLEF 2009 Workshop*, Corfu, Greece.

Comas P, Turmo J and Surdeanu M 2007 Robust question answering for speech transcripts using minimal syntactic analysis *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Comas PR and Turmo J 2008b Spoken document retrieval based on approximated sequence alignment *TSD '08: Proceedings of the 11th international conference on Text,Speech and Dialogue*, pp. 285–292. Springer-Verlag.

Dang HT, Lin J and Kelly D 2006 Overview of the trec 2006 question answering track *Text Retrieval Conference TREC-15*, pp. 99–116, Gaithersburg, MD, USA.

Dang HT, Lin J and Kelly D 2007 Overview of the trec 2007 question answering track *Text Retrieval Conference TREC-15*, Gaithersburg, MD, USA.

Dchelotte D, Schwenk H, Adda G and Gauvain JL 2007 Improved machine translation of speech-to-text outputs *Interspeech'07*, Antwerp. Belgium.

Ferres D, Kanaan S, Gonzales E, Ageno A, Rodriguez H, Surdeanu M and turmo J 2004 Talp-qa system at trec 2004: Structural and hierarchical relaxing of semantic constraints *Text Retrieval Conference TREC-13*, Gaithersburg, MD, USA.

Forner P, Peas A, Alegria I, Forascu C, Moreau N, Osenova P, Prokopidis P, Rocha P, Sacaleanu B, Sutcliffe R and Sang ETK 2008 Overview of the clef 2008 multilingual question answering track *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark.

Fukumoto J, Kato T and Masui F 2002 Question answering challenge (qac-1): Question answering evaluation at ntcir workshop 3 *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan.

Fukumoto J, Kato T and Masui F 2004 Question answering challenge for five ranked answers and list answers - overview of ntcir4 qac2 subtask 1 and 2 *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, Tokyo, Japan.

Fukumoto J, Kato T, Masui F and Mori T 2007 An overview of the 4th question answering challenge (qac-4) at ntcir workshop 6 *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan.

Galliano S, Geoffrois E, Gravier G, Bonastre J, Mostefa D and Choukri K 2006 Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news *Proceedings of LREC'06*, Genoa.

Gauvain J 2006 Overview of the ASR Evaluation *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain.

Giampiccolo D, Forner P, Peas A, Ayache C, Cristea D, Jijkoun V, Osenova P, Rocha P, Sacaleanu B and Sutcliffe R 2007 Overview of the clef 2007 multilingual question answering track *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Gillard L, Sitbon L, Blaudez E, Bellot P and El-Bze M 2006 The lia at qa@clef-2006 *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain.

Gonzalez-Ferreras C, Cardenoso-Payo V and Arnal ES 2008 Experiments in speech driven question-answering *Spoken Language Technology*.

Grau B, Ligozat AL, Robba I, Sialeu M and Vilnat A 2005 Term translation validation by retrieving bi-terms *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria.

Hacioglu K and Ward W 2003 Question classification with support vector machines and error correcting codes *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 28–30. Association for Computational Linguistics, Morristown, NJ, USA.

Harabagiu SM, Moldovan DI and Picone J 2002 Open-domain voice-activated question answering *COLING*.

Hickl A, Williams J, Bensley J, Roberts K, Shi Y and Rink B 2006 Question answering with lcc's chaucer at trec 2006 *The 15th TREC Conference (TREC 2006)*.

Hori C, Hori T, Tsukada H, Isozaki H, Sasaki Y and Maeda E 2003 Spoken interactive odqa system: Spiqa *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pp. 153–156. Association for Computational Linguistics, Sapporo, Japan.

Ittycheriah A and Roukos S 2002 Ibm's statistical question-answering system - trec-11 *Proceedings of the TREC 2002 Conference*.

Jones D, Wolf F, Gibson E, Williams E, Fedorenko E, Reynolds D and Zissman M 2003 Measuring the Readability of Automatic Speech-to-Text Transcripts *Proceedings of Eurospeech'03*, pp. 1586–1588, Geneva, Switzerland.

Kato T, Fukumoto J and Masui F 2004 Question answering challenge for information access dialogue - overview of ntcir4 qac2 subtask 3 *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, Tokyo, Japan.

Kato T, Fukumoto J and Masui F 2005 An overview of ntcir-5 qac3 *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan.

Krsten J, Kundisch H and Eibl M 2008 Qa extension for xtrieval: Contribution to the qast track *Working Notes of CLEF 2008 Workshop*, Aarhus, Denmark.

Kubala F, Schwartz R, Stone R and Weischedel R 1998 Named Entity Extraction from Speech *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Va.

Laurent D, Sgula P and Ngre S 2006 Cross lingual question answering using qristal for clef 2006 *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain.

Lee Y, Al-Onaizan Y, Papineni K and Roukos S 2006 IBM Spoken Language Translation System *TC-STAR Workshop on Speech-to-Speech Translation*, pp. 13–18, Barcelona, Spain.

Li X and Roth D 2002 Learning question classifiers *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7. Association for Computational Linguistics, Morristown, NJ, USA.

Ligozat AL 2006 *Exploitation et fusion de connaissances locales pour la recherche d'informations prcises* PhD thesis Universit Paris-Sud 11 Orsay, France.

Liu Y, Shriberg E, Stolcke A, Hillard D, Ostendorf M and Harper M 2006 Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 1526–1540.

Magnini B, Giampiccolo D, Forner P, Ayache C, Osenova P, Peas A, Jijkoun V, Sacaleanu B, Rocha P and Sutcliffe R 2006 Overview of the clef 2006 multilingual question answering track *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain.

Magnini B, Romagnoli S, Vallin A, Herrera J, Peas A, Peinado V, Verdejo F and de Rijke M 2003 The multiple language question answering track at clef 2003 *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway.

Magnini B, Vallin A, Ayache C, Erbach G, Peas A, de Rijke M, Rocha P, Simov K and Sutcliffe R 2004 Overview of the clef 2004 multilingual question answering track *Working Notes for the CLEF 2004 Workshop*, Bath, UK.

Miller J and Weinert R 1998 *Spontaneous Spoken Language. Syntax and Discourse*.

Mitamura T, Nyberg E, Shima H, Kato T, Mori T, Lin CY, Song R, Lin CJ, Sakai T, Ji D and Kando N 2008 Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access *Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan.

Moldovan D, Harabagiu A, Girju R, Morarescu P, Lacatusu F, Novischi A, Badulescu A and Bolohan O 2002 Lcc tools for question answering *Proceedings of the 2002 Text Retrieval Conference*.

Molla D, Cassidy S and van Zaanen M 2007 Answerfinder at qast 2007: Named entity recognition for qa on speech transcripts *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Molla D, van Zaanen M and Pizzato L 2006 Answerfinder at trec 2006 *The 15th TREC Conference (TREC 2006) proceedings*.

Monz C and de Rijke M 2001 The university of amsterdam's textual question answering system *Text Retrieval Conference TREC-10*, Gaithersburg, MD, USA.

Neumann G and Wang R 2007 Dfki-lt at qast 2007: Adapting qa components to mine answers in speech transcripts *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Odell M and Russell R 1918 Us patent numbers 1261167 (1918) and 1435663 (1922) U.S. Patent Office.

Pardio M, Gmez J, Llorens H, Muoz-Terol R, Navarro-Colorado B, Saquete E, Martnez-Barco P, Moreda P and Palomar M 2008 Adapting ibqas to work with text transcriptions in qast task: Ibqast *Working Notes of CLEF 2008 Workshop*, Aarhus, Denmark.

Peas A, Forner P, Sutcliffe R, Rodrigo A, Forascu C, Alegria I, Giampiccolo D, Moreau N and Osenova P 2008 Overview of the respubliqa 2009: Question-asnwering evaluation over european legislation *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark.

Reyes-Barragan A, Villasenor-Pineda L and y Gomez MM 2009 Inaoe at qast 2009: Evaluating the usefulness of a phonetic codification of transcriptions *Working Notes of CLEF 2009 Workshop*, Corfu, Greece.

Rosset S, Galibert O, Adda G and Bilinski E 2007 The limsi participation to the qast track *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Rosset S, Galibert O, Bernard G, Bilinski E and Adda G 2008 The limsi participation to the qast track *Working Notes of CLEF 2008 Workshop*, Aarhus, Denmark.

Sasaki Y, Chen HH, hua Chen K and Lin CJ 2005 Overview of the ntcir-5 cross-lingual question answering task (clqa1) *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan.

Sasaki Y, Lin CJ, hua Chen K and Chen HH 2007 Overview of the ntcir-6 cross-lingual question answering (clqa) task *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan.

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

Toney D, Rosset S, Max A, Galibert O and Bilinski E 2008 An evaluation of spoken and textual interaction in the ritel interactive question answering system In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (ed. (ELRA) ELRA), Marrakech, Morocco.

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

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

Turmo J, Comas P, Rosset S, Galibert O, Moreau N, Mostefa D, Rosso P and Buscaldi D 2009 Overview of qast 2009 *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece.

Turmo J, Comas P, Rosset S, Lamel L, Moreau N and Mostefa D 2008 Overview of qast 2008 *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark.

Vallin A, Giampiccolo D, Aunimo L, Ayache C, Osenova P, Peas A, de Rijke M, Sacaleanu B, Santos D and Sutcliffe R 2005 Overview of the clef 2005 multilingual question answering track *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria.

van Schooten B, Rosset S, Galibert O, Max A, op den Akker R and Illouz G 2007 Handling speech input in the ritel qa dialogue system *in InterSpeech'07*, Anvers, Belgique.

Voorhees EM 2000 Overview of the trec-9 question answering track *Text Retrieval Conference TREC-9*, pp. 71–80, Gaithersburg, MD, USA.

Voorhees EM 2002 Overview of the trec 2002 question answering track *Text Retrieval Conference TREC-11*, Gaithersburg, MD, USA.

Voorhees EM 2003 Overview of the trec 2003 question answering track *Text Retrieval Conference TREC-12*, pp. 54–68, Gaithersburg, MD, USA.
Voorhees EM 2004 Overview of the trec 2004 question answering track *Text Retrieval Conference TREC-13*, Gaithersburg, MD, USA.
Voorhees EM and Dang HT 2005 Overview of the trec 2005 question answering track *Text Retrieval Conference TREC-14*, Gaithersburg, MD, USA.
Voorhees EM and Tice DM 1999 The trec-8 question answering track report *Text Retrieval Conference TREC-8*, pp. 77–82, Gaithersburg, MD, USA.
Voorhees EM and Tice DM 2001 Overview of the trec 2001 question answering track *Text Retrieval Conference TREC-10*, pp. 42–51, Gaithersburg, MD, USA.
Whittaker E, Chatain P, Furui S and Klakow D 2005 Trec2005 question answering experiments at tokyo institute of technology *Proceedings of the 14th Text Retrieval Conference*.
Whittaker E, Novak J, Heie M and Furui S 2007 Clef2007 question answering experiments at tokyo institute of technology *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.