Integrating Spoken Dialog and Question Answering: the Ritel Project

Sophie Rosset (1), *Olivier Galibert* (1)

(1) Human-Machine Communication Dept.,
LIMSI-CNRS, Orsay, France
{rosset, galibert}@limsi.fr

Gabriel Illouz (1)(2), *Aurélien Max* (1)(2)

(2) Université Paris Sud 11
Orsay, France
{gabrieli, amax}@limsi.fr

Abstract

The Ritel project aims at integrating spoken language dialog and open-domain information retrieval to allow a human to ask general questions (e.g. *Who is currently presiding the French Senate?*) and refine her search interactively. This project is at the junction of several distinct research communities, and has therefore several challenges to tackle: real-time streamed speech recognition with very large vocabulary, open-domain dialog management, fast information retrieval from text, query cobuilding, communication between information retrieval and dialog components, and generation of natural sounding answers. In this paper, we present our work on the different components of Ritel, and provide initial results.

Index Terms: spoken dialog system, information extraction, question-answering

1. Introduction

Searching for information can be done using one of two main paradigms: document retrieval and information extraction. In the former, documents matching a user query, usually a few keywords, are returned. Based on the assumption that the theme of these documents is the one that is best described by the query, they constitute a pool in which the user might find information that can meet some need. This need can be very specific (e.g. *Who is currently presiding the French Senate?*), or it can be theme-oriented (e.g. *I'd like information about the French Senate*). The other approach to search is embodied in so-called question answering systems, which given a specific spelled out question return the most probable answer (e.g. *Who won the 2005 Tour de France? Lance Armstrong.*)

The name **Dialog System** covers a large domain, but usually denotes a system enabling interaction between humans and computers in a restricted field of knowledge [1]. Over the last few years [2], this definition has started to widen to allow for a larger skillset on the computer side, especially with progress on question answering. Both **document retrieval** and **question answering** are active fields of research whose main limitations are mostly due to their inability to interpret language. While our ongoing projects include research on automatic methods to match documents to queries and answers to questions, we consider that search as a computer-assisted task is a very promising basis for new advances in human-computer interaction. Until recently, there were very few projects on interactive open-domain information retrieval (e.g. [3]), and it now seems there exists a growing interest with projects along the same lines as the project we describe in this paper ([4, 5]).

RITEL (standing for *Recherche d'Information par Téléphone*) is an ongoing project at LIMSI concerned with studying human-computer interaction in the context of dialogs for information retrieval in unrestricted domains. The platform we are building integrates speech recognition, utterance analysis, information retrieval, natural language generation and speech synthesis. This paper presents the current state of the system and of its components. Section 2 presents an overview of the RITEL platform. *Speech activity detection and recognition* are briefly described in section 3. Section 4 presents the *non-contextual analysis* on which all the following components (information extraction, dialogic interactions...) are based. Section 5 presents the *information retrieval* component, and the *natural language generation* component is presented in section 6. We conclude with some perspectives in section 7.

2. The RITEL platform

An overview of the spoken language system architecture is shown in Figure 1. The main components we have implemented are the Automatic Speech Recognizer, the Non-Contextual Analyser, Information Retrieval and Natural Language Generation.[1]. All these components are completely open to each other, communicating through a message-passing infrastructure, which leads to a distributed dialog management.

3. Speech activity detection (SAD) and recognition (ASR)

The SAD system is identical to the one in [6], and the ASR system is similar with new specialized acoustic models which have been built on 4 hours of audio data and a slightly larger corpus for the language models. The vocabulary is composed of about 65K words. The out-of-vocabulary rate is 1.3% on the development corpus and 1.7% on the test corpus, which is reasonable for an open-domain system for French. Performance is 0.5RT streamed for a word-error-rate of around 28%. Efforts are made to increase the amount of data which should lead to noticeable improvements on the quality of the recognition.

The target speed performance of 0.85 real time has been met. At that level, the ASR terminates when the user stops speaking, thus enabling extremely snappy and natural interaction.

[1] The Text-To-Speech synthetizer is a commercial product which will not be described here

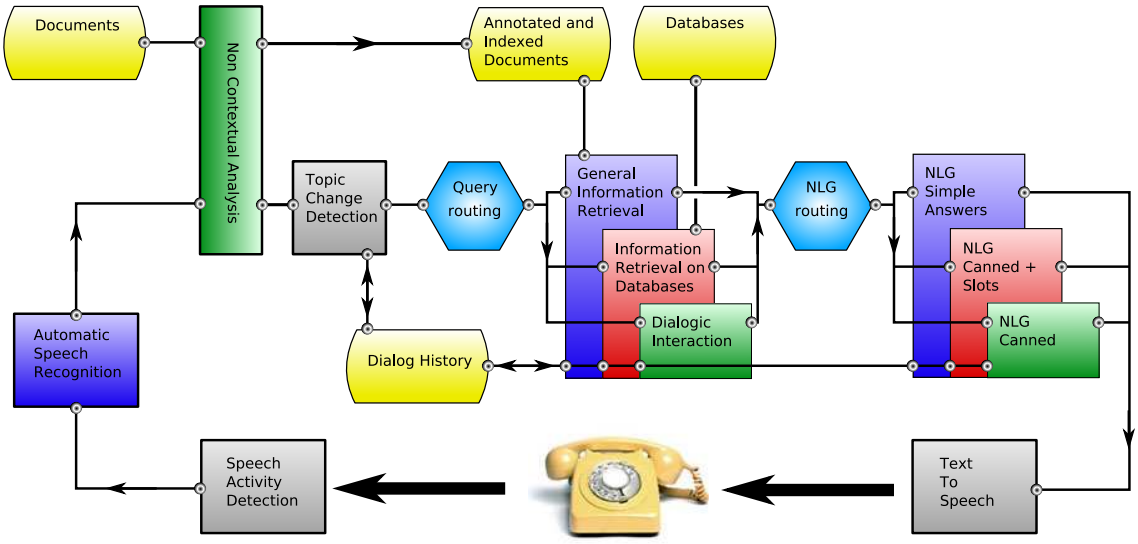

Figure 1: Overview of the RITEL system

English data:

_prep **in** _org **NIST** _NN **metadata evaluations**
_action **reported** _NN **speaker tracking**
_score **error rates** _aux **are** _prep **about**
_val_score **15 %**

French data:

_org **Airbus** _aux **a** _action **vendu** _val **10**
_obj_val _prod **A380** _prep _org **Fedex**

Figure 2: Examples of Pertinent Information from a complete sentence of our English and French data collection

4. Non-contextual analysis (NCA)

We call our analysis *non-contextual* because no dialog information or previous sentences are used. The general objective of the NCA is to detect the pertinent information of a sentence (a user question or utterance, spoken or written). Figure 2 shows examples of what we call *pertinent information*, which can be of different categories: named entities, linguistic entities (actions, prepositions), specific entities (scores, val_scores). Moreover, we consider that all remaining words have potentially important information and should thus be also annotated. The general idea for them is to try and annotate the longest groups with a unique, coherent meaning. Also, RITEL being designed for use in an spoken interactive question-answering system, it has two important constraints: work on both written and spoken data (text and queries), and be as fast as possible.

4.1. Definition of entities

Different entities at different linguistic levels have been defined. We extended the definition for the named entities to *expression describing one specific element of a given kind*, where "kind" covers location, person, organization, etc. For instance *president of the United States in 2006* is a person named entity by our definition. Additionally, we decided to allow for hierarchical named entities, so the previous example would also have *United States* tagged as organization and *2006* as date. These entities are very useful for the QA system as search keys, answer chunks and for question typing. For example, *NIST* is an ORG, the *2006 Cannes festival* an EVENT and *veni vidi vici* is a CITATION. Table 1 gives an overview of the different types of entities tagged by our system.

4.2. System and preliminary results

The system is currently rule-based. It needs to be able to manage lists for initial detections, local context and easy categorizations. Given these constrains, a word-based regular expression engine was implemented with some added NLP-specific features such as *positive and negative lookaheads*, *named classes and macros*, *shy and greedy groupings*, *strategies for prioritizing rule application*, and *external word categorisation*. The analyzer uses various lists which contain about 2 600 names, 500 countries, 185 000 cities and 300 languages. The analysis is done in 40 steps and takes a handful of milliseconds per sentence.

An evaluation of named and extended entities detection gave a F-measure from 82% on broadcast news to 88% on spoken queries.

5. Interaction and information retrieval

As shown in Figure 1, there is no specific dialog manager. *Dialog management* is distributed over all aspects of the platform, thus the approach is completely integrated. *Topic change detection*, *Query*

Named entities	_org NIST _eve 2006 Cannes festival who said _cit veni vidi vici
Non-specific entities	_Eve Cannes festival the _Pers president said ...
Multi-level extended entities	Functions, titles (president, bishop...) Colors, animals...
Hierarchical superclassing	bishop → religious function → function
Topic markers	I'm interested in _litterature novels by ... won the _sport Mundial in 1998
Question markers	_Qwho who wrote that book _Qmeasure how many hours of transcription do you need
Interaction markers	_DA_close goodbye _DA_yes yes please
Information chunks Compound nouns Action verbs and composites Linguistic entities	 the _NN local farmers are ... the _NN multiracial elections he _action won the they _action gave up _adj_comp highest prize it happens _adv often when ...

Table 1: Summary of the Typed entities

routing and *NLG routing* can all be seen as part of dialog management.

The general idea of *Topic change detection* is to answer to the following question: *Do we want to complete the request with elements from the previous exchanges?* This module uses the topics markers to detect the topic of the user utterance. On topic change the history is flushed.

After *Topic change detection*, the user utterance is routed. Using the question and its analysis, and the interaction markers, the *Query routing* component classify the type of utterance:

- general information retrieval
- information retrieval on databases
- pure dialogic interaction or unclassifiable utterance

5.1. General information retrieval

The *General information retrieval* component corresponds to a classical *Question-Answering* system. It handles search in documents of any types (news articles, web documents, transcripted broadcast news, etc.). For speed reasons, the documents are all available locally and preprocessed: they are first normalized, and then analyzed with the NCA module described in section 4. The (type, values) pairs are then managed by a specialised indexer for quick search and retrieval. This somewhat bag-of-typed-words system works in six steps:

1. Detection of the answer type using Question Marker, Named, Non-specific and Extended Entities co-occurences.

- _Qwho → _pers or _pers_def or _org
  - who sold Manhattan
- _Qwhat + _function → _pers
  - what is the name of the pope

2. Decision on the relative importance of the different entities in the utterance.

- Named entity > NN > adj_comp > action > subs ...

3. Document query creation, first with all the entities and then a series of currently handcrafted backoff queries relaxing some of the constraints:

- Increase the allowable distance between entities.
- Allow type changes for some entities (e.g. _loc → _org).
- Allow including or included values in the document (e.g. Bush → Georges Bush).
- Drop some of the entities using their relative importance.

4. Passage retrieval: sending the queries to the indexation server and getting document snippets (sentence or groups of sentences) back.

5. Candidates extraction: finding the entities in the snippets with the expected type of the answer.

6. Answer selection: a clustering of the candidate answers is done, based on frequencies. For now, the most frequent answer wins, and the distribution of the counts gives an idea of the confidence of the system in the answer.

This *baseline* system fits the time constraint. Preliminary results on CLEF'04 test set gave us a MRR of 60% for factual questions.

5.2. Information retrieval on databases

Sometimes, if accurate structured information is available (local databases, TV schedules, IMDB, etc.), it is more efficient to use specific handling. This is similar to what is done in traditional limited-domain dialogue systems: first pick the latest query type in the dialogue history if one is not present in the user utterance, then complete the query slots needed for the query type from the dialogue history, and then lookup in the database. The next versions should also complete the answer with results from General Information Retrieval.

5.3. Dialogic interaction

Except for the information retrieval interaction, the global system has to deal with **pure dialogic interaction**. This concerns mainly the management of general interaction (such as *please repeat*, *I don't understand*, *goodbye*), and dealing with unclassified user utterances which are mostly utterances with too many errors and which trigger a non-understanding reaction from the system. Some of the reactions are: *send the guide*, *send a goodbye sebtence then hang up*, and *reformulate or repeat*.

6. Natural language generation of answers

The generation of natural answers in question answering systems has not received a lot of attention until recently, as it is traditionally not part of what is evaluated. In the context of RITEL, it is however extremely important to make the user feel that the system is *cooperative* by means of its answers. This can be achieved by helping the user to iteratively refine her query, suggesting possible areas of interest, or completing the answers when appropriate (e.g. [7]). Moreover, care must be taken so that the output of the system sound natural: avoid redundancy, indicate system confidence using language, take the history of the dialog into account, etc. Therefore, a close interaction with the information retrieval component seems essential.
We have implemented a baseline for the generation of answers that will be used as back-off in case more advanced techniques cannot be applied. Dialogic interaction can often be dealt with using canned text for which several variants can be used (e.g. *Could you please repeat your question?*). For questions for which an answer pattern can be associated, templates with canned text and slots are used (e.g. *The <function> of <country> is <answer>*). Because handcrafting and maintaining templates is expensive, this is only practical for recurrent question patterns. Producing the answer accompanied by the number of times it was found in documents is used otherwise (e.g. *Neil Armstrong (400 documents)*).

The time constraint on the interaction discourages the use of deep NLG techniques, which would require a very fine-grained analysis of the question. However, some elements from the question such as syntactic structure, concept lexicalizations and expected answer type should be reused [8]. Three important aspects are to use formulations that indicate the system's confidence in the extracted answers (e.g. *According to the White House webpage, the US president in 1943 was Franklin D. Roosevelt*), to allow for implicit confirmation when the system is unsure of the question (e.g. *The national anthem of the Netherlands is...*), and to help the user to refine her search when an ambiguous search yields several possible answers (e.g. *Several Eiffel towers exist in the world. Are you interested in the one in Paris or in another one?*).

7. Conclusion and Perspectives

In this paper, we presented the RITEL platform. The dialog system deals succesfully with a major constraint: reaction time. The system is snappy and answers are instantaneous. Most of the work presented in this paper is corpus-based (notably, non-contextual analysis and general information retrieval). *Non-contextual analysis*, which is a *high-level light analysis*, works on both spoken (user utterances) and written (documents with answers) data. It is used to detect named and specific entities and information chunks. Initial experiments showed a F-measure from 82% on broadcast news to 88% on spoken queries. *Information retrieval* is done using both a general system and a specialized system which uses specific structured data. The general information retrieval component, which uses the NCA component, handles open-domain search in large sets of documents. Preliminary results on the Clef'04 test set gave us a MRR of 60% for factual questions. Information retrieval, dialogic interactions handling, context management and natural language generation are completely integrated making dialogue management fully distributed.

Our work on general information retrieval is ongoing in several directions: automatic leaning of query lists, typed patterns-based candidate extraction for better extraction quality, and pattern scoring for better clustering.

Regarding dialogic interaction, the next version of the system should be able to answer questions on the identity of the sources and cite complete document snippets if requested by the user.

We are currently investigating two main areas in improving the answers generated by the NLG component. First, we want to make IR and generation work in parallel and hand-in-hand to produce stalling, but natural sounding answers while IR is running more complex algorithms using backchannel and formulations that postpone the actual answer. Second, we are interested in reusing existing formulations found in documents and completing them with additional information. This is motivated by the fact that answers from the system will often bring related questions from the user (this is sometimes referred to as "berry picking"). A possible approach is to align at some level text spans containing an answer [9], merge those alignments and decide which frequently cooccurring elements might be of interest before producing the answer. For instance, *When was Henry IV assassinated?*[2] could be answered concisely by *Henry IV was assassinated in 1610*, but an answer such as *Henry IV was stabbed to death in 1610 by Ravaillac* would further provide the method and murderer.

8. References

[1] J. R. Glass, J. Polifroni, S. Seneff, and V. Zue. Data collection and performance evaluation of spoken dialogue systems: the mit experience. In *in ICSLP'00*, Pekin, Chine, 2000.

[2] J. R. Glass. Challenges for spoken dialogue systems. In *in ASRU'99*, Keystone, Colorado, 1999.

[3] Sanda Harabagiu, Dan Moldovan, and Joe Picone. Open-domain voice-activated question answering. In *in Proceedings of International Conference On Computational Linguistics*, 2002.

[4] Rieks op den Akker, Harry Bunt, Simon Keizer, and Boris van Schooten. From question answering to spoken dialogue: Towards an information search assistant for interactive multimodal information extraction. In *InterSpeech*, Lisbon, September 2005.

[5] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, and S. Furui. Deriving disambiguous queries in a spoken interactive odqa system. In *in ICASSP'03*, Hong Kong, 2003.

[6] O. Galibert, G. Illouz, and S. Rosset. Ritel: An open-domain, human-computer dialog system. In *in InterSpeech'05*, Lisbon, Portugal, 2005.

[7] W. Bosma. Extending answers using discourse structure. In *in RANLP Workshop on Crossing Barriers in Text Summarization Research*, Borovets, Bulgaria, 2005.

[8] S. Mendes and V. Moriceau. L'analyse des questions : intrts pour la gnration des rponses. In *In Proceedings of the Workshop on Question Answering, TALN'04*, Fz, Maroc, 2004.

[9] E. Marsi and E. Krahmer. Explorations in sentence fusion. In *in Proceedings of the 10th European Workshop on Natural Language Generation, Aberdeen*, Aberdeen, Scotland, 2005.

[2] We assume that no previous question was related to Henry IV.