Ritel: An Open-Domain, Human-Computer Dialog System

Olivier Galibert, Gabriel Illouz, Sophie Rosset

Spoken Language Processing Group, LIMSI-CNRS,
B.P. 133, 91403 Orsay cedex, France
{galibert, gabrieli, rosset}@limsi.fr
http://www.limsi.fr

Abstract

The project RITEL aims at integrating a spoken language dialog system and an open-domain question answering system to allow a human to ask general questions ("Who is currently presiding the Senate?") and refine the search interactively.

As this point in time the RITEL platform is being used to collect a human-computer dialog corpus. The user can receive factual answers to some questions (**Q :** who is the president of France, **R :** Jacques Chirac is the president for France since may 1995). This paper briefly presents the current system, the collected corpus, the problems encountered by such a system and our first answers to these problems.

When the system is more advance, it will allow measuring the net worth of integrating a dialog system into a QA system. Does allowing such a dialog really enables to reach faster and more precisely the "right" answer to a question?

1. Introduction

Recent progress in the last years in both Automatic Speech Recognition (ASR) and Question Answering (QA) systems opens the way to a new generation of dialog systems allowing a human user to, through a phone or in front of a screen, ask a computer for information on any subject. Building such a system is the aim of the RITEL project. The human user shall be able to ask a general question ("Who is currently presiding the Senate?") and refine the search interactively.

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 [4]. Over the last few years [3] this definition has started to widen to allow for a larger skillset on the computer side, especially with progress on QA.

Generally speaking, a dialog is a succession of interactions between speakers in a given context. A human-computer dialog system analyses and acts on the user requests as a function of the task at hand, the previous interactions and the user behaviour. Its aim is to provide the user with the information they're looking for while keeping a smooth and natural interaction flow. Today, most dialog systems work in restricted domains such as time information on transportation facilities (trains, planes) or tourist information. Some examples are the European project Le3-Arise (train reservation), the American projects ATIS and DARPA Communicator (plane trips) and the French project MEDIA (tourist informations). These projects included quantitative evaluations and proved the feasibility of such systems by in particular pushing models for dynamic dialog management and adapted natural language generation. A dialog system uses varied and complex knowledge sources: acoustics, phonetics, lexical information, morphology, syntax, semantics, pragmatics, and knowledge about dialog in general, the task and user behaviour. Specialized modules taking charge of specific computations rely on one or more of these sources. Examples of these modules are the ASR (acoustics, phonetics, lexical information), utterance analyser (syntax, semantics), dialog manager (pragmatics, task, user behaviour...).

In the Information Retrieval (IR) and Question Answering (QA) domains progress has also been pushed through evaluation campaigns (USA: TREC 1998-2003, Europe: CLEF, France: Equer/Technolangue 2004). These systems only handle independent questions and provide one answer for each. Some trials have been done on successions of questions on a similar topic, but these weren't dialogs: no real interaction nor actual negotiations possible.

RITEL aims at the integration of these and as such is rich and complex and has to face a number of obstacles. Some are obvious: the speech recognition which has to be large vocabulary and open-domain with a hard real-time constraint, the dialog management which also has to be open-domain, the information exchange between QA and dialog, the generation of answers combining information from multiple documents.

Our aim is to build a research platforms which will allow us to try and solve these problems and any other that is encountered along the way. This paper presents the current state of the architecture of the system and of the modules which are built in it. Then we briefly present the speech corpus that is being collected and annotated. And finally we give some perspectives on this study.

2. The RITEL system

An overview of the spoken language system architecture is shown in Figure 1. The main components for spoken language are the speech recognizer, the entities tagger, the dialog manager, the question-answering system, the natural language generator system and the Text-To-Speech synthetizer. They communicate through a message-passing infrastructure.

The dialog manager controls and organise the interaction. It manages the entities tagger and the information passed through the QA system.

2.1. Speech activity detection and recognizer

The Speech Activity Detection (SAD) component is derived from LIMSI's Conversational Telephonic Speech (CTS, also known as Hub 5) segmenter: GMMs are used to model speech and silence. A standard streamed Viterbi beam decoding then does the segmentation, outputting results as the beam closes. This gives a good detection quality with a mean detection de-

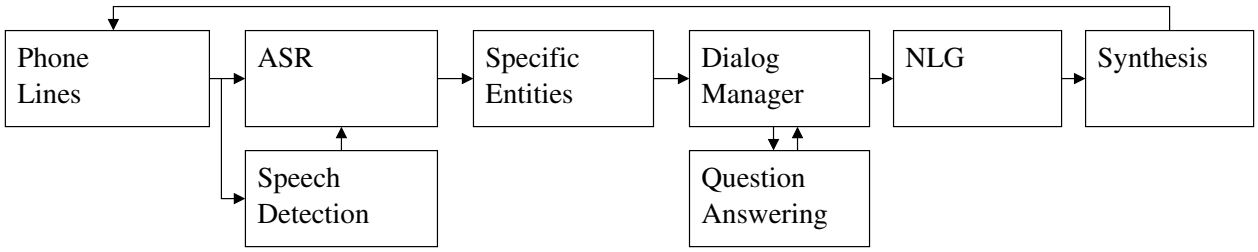

Figure 1: Overview of the RITEL system

lay of around 0.4s with some false positives on loud breaths in particular but no false negatives to speak of.

The ASR system is a single-pass lattice decoder [2] using triphone acoustic models and a trigram language model followed by a quadrigram consensus rescoring. The acoustic models are the ones created for the dialog system Arise [6].

The language models have been created by interpolating various sources on a subset of the collected corpus (section 3) slated as dev. These sources are:

- RITEL training corpus
- Newspapers text from 1988 to today
- Manual transcriptions of radio and TV news shows
- Questions / answers from the Internet (f.i. Madwin.fr)

The vocabulary is composed of about 65K words. It was chosen as the 65K most probable words in a unigram language model created by interpolation with the same method and sources than the decoder language models. It has an Out Of Vocabulary rate of 1.3% on the dev and 1.7% on the test, which is reasonable for an open domain system in French.

The next versions of these models will integrate CTS transcriptions (giving disfluencies and spontaneity in general) and also transcriptions from Arise (giving questions, negations, general dialog interactions). Once enough data is collected specialized acoustic models will also be estimated.

2.2. Specific entities detection

The specific entities detection system takes an utterance as input and outputs the same utterance with the interesting parts typed and tagged. This step is also often called non-contextual utterance analysis, *context* meaning here the previous requests and answers exchanged by the user and the system. The specific entities are from 3 categories:

- Named entities like people (<pers>), products, titles and commercial names (<prod>), time markers (<time>), organizations (<org>), lexical units (<ul>) and places (<loc>).
- Syntaxic request markers (who, where, when, how, how much...), which are rewritten into class markers (Twho, Twhere, Twhen, Tamount...).
- Semantic request markers, giving topics (literature, geography, history, social, theater, politics, economy...) and subtopics (author, president, spelling, population count...).

All these entities are extracted using a set of rules which take the form of regular expressions on words. Macros (local sub-rules) and classes (lists) can be used in the rules definitions. The analysis steps are split depending on their function:

- The first step rewrites numbers:
  nineteen hundred eighty two
  becomes
  <1982> nineteen hundred eighty two </1982>.
- The second step is a lexical analysis. It normalizes the syntaxic markers of requests:
  I'd like to know the name ...
  becomes
  I'd like <Twho> to know the name </Twho> ....
- The third step annotates topics, subtopics and named entities:
  <Twho> who </Twho> wrote red and black
  becomes
  <subtopic> <Tauthor> <Twho> who </Twho> wrote </Tauthor> </subtopic> <prod> Red and Black </prod>.

2.3. Dialog Manager

The first version of the dialog manager was designed to incite people to talk as much as possible, reformulate their request in as many ways as possible, to refine their question while keeping a reasonably natural interaction. That dialog manager can hence be considered an ELIZA variation [9]. The second version allows to search information in databases as appropriate. The implementation is largely based on the same engine as LIMSI's ARISE [6].

The current dialog manager role is to:

- analyze in context (i.e. taking into account what was said before) the semantic frame the specific entities detector outputs
- extract the relevant information to search in the databases
- build semantic frames for the natural language generation system
- pick a natural language generation strategy

A dialog is divided into three phases around which all interactions take place: the actual *Information Search* phase preceded by some *Opening formalities*, which may include requesting a general usage information message, and optionally closed by some *Closing formalities*. The opening formalities, for most of the users, share the first utterance with the initial formulation of the request.

Each semantic frame as sent by the specific entities detector is first analyzed in context, i.e. taking into account the history of the interaction. The new, in-context frame is sent to the decision module which rewrites it again, this time using both a dialog model (how interactions go in general, whatever the subject) and a task model (how specifically requesting specific information and refining the request happens). If according to these models the current request is of the kind which can be answered factually by searching in one of the available databases, the search module extract the relevant keys and does the search. Otherwise the incitation module isolates the topic of the request in order to generate an answer which, while not actually answering the question, shows the system has understood something and urges the user to refine or reformulate his question. These

two modules generate new semantic frames that are sent to the natural language generation module.

All the operations of these modules are written using a subset of order-2 formal logic using a declarative form.

2.4. Question Answering system

Current searches can only be done in fixed databases, but a full-blown QA system is in the process of being connected to the dialog manager. We won't describe the usual architecture of QA systems here, the interested reader can for instance read [5]. Some interesting points are already appearing though which are to be taken into account:

- Speed is critical, the user will hang up if the system stops talking for too long and traditional QA systems answer in a matter of minutes or worse hours.
- QA systems do their own analyses of the request, in particular where it comes to named/specific entities, and while similar to the ones done for the dialog manager they are far from identical. Some impedance matching work has to be done to make the bidirectional communication workable.
- The user gives feedback on the quality of the answers it gets and can hence steer the QA system towards the information he actually wants. Current systems, being non-interactive, have to use *Blind Relevance Feedback* approaches to increase the quality of their answers, which is pretty much the opposite.

2.5. Natural Language Generation and Synthesis

Part of the generation is currently done by the dialog manager, which generates a semantic frame sent to the NLG module. Two strategies are followed, depending on whether the search module or the incitation module is activated:

- The request is answerable, the search module kicks in. If enough information is available for searching the appropriate databases it is done and a semantic frame with the useful elements of the question and the search result is generated. Otherwise a frame with the list of missing elements is made.
- The system is not able to answer the request, the incitation module takes over. It generates a frame with the elements we have understood in the request so that the NLG module can make a sentence urging the user to speak more, refining or explaining his request.

The NLG module itself is based on a set of template sentences. They allow to answer at different levels depending on the dialog state and its history as sent by the dialog manager. It can for instance ask the user to refine his request, or it can explain that it knows the search topic but fails to understand the link with the actual request object, or even tell the user that the system utterly failed to understand what was asked of it and please try starting from scratch.

The audio synthesis is a simple in-house diphone concatenation system of medium quality. We are going to replace it with a commercial system closer to the state-of-the-art which will help getting more natural interactions with the human user.

3. Corpus

The corpus was collected between September 2004 and February 2005. 13 persons called the *Ritel* system. Each subject had received a list of 300 possible questions. They were told to feel free to ask the system what they want however they want. Table 1 show a summary of the data collected. Of the 6 hours of user speech one hour has been set aside for development (dev) and one hour for testing (test) of the modules, especially the speech recognition and the specific entities detection.

# dial.	duration of user speech	# user's queries	# user's words
582	6h	5360	60k

# distinct user's words	# user's queries per dial.	mean duration of user's speech/dial
2876	9	33s.

# Topics & # Sub-Topics	# Ling. Ent.	# Nam. Ent.
1120	6860	7000

Table 1: Summary of the RITEL corpus

S: Hello and welcome to Limsi's Ritel service. Do you want the help message?
U: err no thanks I'd like the capital of France please
S: The capital of France is Paris. Do you want another information?
U: err who is its president
S: Jacques-Chirac is the president of France. Do you want another information?
U: and err the one for Bulgaria
S: Georgi-Parvanov is the president of Bulgaria. Do you want another information?
U: hm and which country Ion Iliescu is the president then?
S: Ion-Iliescu is the president of Romania. Do you want another information?
U: since when
S: Ion-Iliescu is the president of Romania since december 2000. Do you want another information?
U: none thanks goodbye
S: Thanks for having used Limsi's Ritel service. We thank you for you call. Goodbye.

Figure 2: Translated sample of dialog collected with factual answers

Figure 2 shows a sample of dialog with factual answers generated from the results the search module of the dialog manager. Conversely, figure 3 shows a sample of dialog with incitations generated by the incitation module and its Natural Language Generation counterpart as described in sections 2.3 and 2.5.

The corpus is structured into topics and sub-topics hierarchically organised. Topics consists of a classical list of domain such as *history, geography, literature, movies, arts, sciences* etc. The sub-topics give additional informations and are the kind of *date of birth, date of death, profession, author, physics, biology* etc. The corpus is under annotating process. Figure 4 gives the repartition of the topics in 53 dialogs (593 user's utterances) recorded between January and March 2005.

4. Conclusions and Perspectives

The aim of the RITEL project is to build a research platform linking a Question Answering system and a Spoken Language Dialog System. Our first step was to show the feasibility of such a project. Right now the system allows a natural, if somewhat limited in scope, interaction between a user looking for information and itself. It already allowed to collect a linguistically rich and realistic corpus. It also showed that some (minimalist)

S: Hello and welcome to Limsi's Ritel service. Do you want the help message?
U: I'd like some information on the ONU
S: You're looking for information about the ONU. Could you be more specific ?
U: which countries are in it
S: Your request is about geography. But more specifically? Are you interested in a specific country? A town?
U: who wrote to be or not to be
S: I do not know who is the author of to be or not to be. I've lost my connection with the information search server. Do you want another information?

Figure 3: Translated sample of dialog collected with incitations

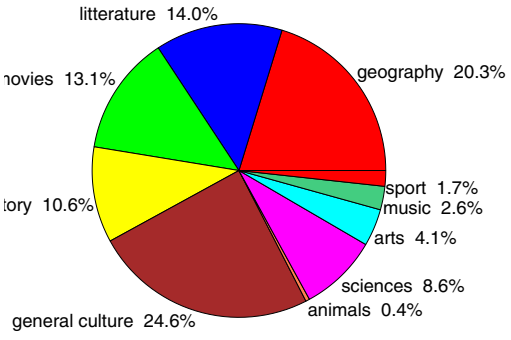

Figure 4: Repartition of topics in a sub-part of the collected data

information research is actually possible, given that some of the dialogs did end in a correct answer. This corpus also gives a preliminary idea of the user behaviour in such a system. This study allowed us to specify along which axes our research and development efforts should go:

- Real-time, streamed Automatic Speech Recognition on large, dynamically-extensible vocabulary. Also open domain and with dynamic adaptation of the models during the dialog.
- Open domain dialog management, and handling of the multi-level information returned by the QA system.
- Information retrieval: ordering and presentation of the answers depending on the dialog state.
- Information types needed to allow the dialog manager and the other modules to rate their analysis and answers according to the dialog state and the QA results.
- Studies on the cost of the various strategies for QA search run control, with startup either continuous or at the request of the dialog manager and the search running while the dialog continues.
- General answer generation, automatic summarization, ...

Another important point of our work will be about evaluating the system. There has been a lot of work on the general problem of evaluating dialog systems. We can in particular cite the Technolangue project MEDIA/EVALDA [7]. For the QA systems evaluation numerous campaigns have also happened, in particular in the series of TREC conferences [1]. In the RITEL project we will have to not only evaluate the dialog manager or the QA system per se, but also the interaction between them. Hence, in addition to the usual criteria and protocols for the evaluation of that kind of systems we will have to define criteria and protocols specific to our kind of platform. They will have to allow measuring the net worth of integrating a dialog system into a QA system. Does allowing such a dialog really enables to reach faster and more precisely the "right" answer to a question?

5. References

[1] Ellen M. Voorhees, "Overview of TREC 2003", In Voorhees and Buckland, 2003.

[2] Jean-Luc Gauvain, Lori Lamel, Holger Schwenk, Gilles Adda, Langzhou Chen, and Fabrice Lefevre. Conversational telephone speech recognition. In Proceedings of ICASSP, pages I-212-215, Hong Kong, April 2003.

[3] Glass J. R., Challenges for spoken dialogue systems, Proceedings of *ASRU'99*, Keystone, Colorado, 1999.

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

[5] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba, A.Vilnat "How NLP Can Improve Question Answering", revue Knowledge Organization, Vol. 29, N3-4, 2002, pages 135-155

[6] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus et P. Morarescu, The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering, Proceedings of *Association for Computational Linguistics*, 2001

[6] L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts, The LIMSI ARISE System, *Speech Communication* Vol. 31(4):339-354, 2000.

[7] Devillers L., Maynard H., Rosset S., Paroubek P., McTait K., Mostefa D., Choukri K., Bousquet C., Charnay L., Vigouroux N., Bchet F., Romary L., Antoine J.Y., Villaneau J., Vergnes M., Goulian, The French MEDIA/EVALDA Project: the Evaluation of the Understanding Capability of Spoken Language Dialogue Systems, *LREC'04*, 2004.

[8] M. Walker, D. Litman, C. Kamm, A. Abeilla , Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies, *Computer Speech and Language*, 1998.

[9] J. Weizenbaum, "ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine", Communications of the ACM Volume 9, Number 1 (January 1966): 36-35.