The Ritel Corpus - An annotated Human-Machine open-domain question answering spoken dialog corpus

Sophie Rosset, Sandra Petel
LIMSI - CNRS

Abstract
In this paper we present a real (as opposed to Wizard-of-Oz) Human-Computer QA-oriented spoken dialog corpus collected with our RITEL platform. This corpus has been orthographically transcribed and annotated in terms of Specific Entities and Topics. Twelve main topics have been chosen. They are refined into 22 sub-topics. The Specific Entities are from five categories and cover Named Entities, linguistic entities, topic-defining entities, general entities and extended entities. The corpus contains 582 dialogs for 6 hours of user speech.

1. Introduction
The Ritel project aims at integrating a spoken language dialog system and an open-domain information retrieval system to allow a human to ask a general question (i.e. “Who is currently presiding the Senate?” or “How did the price of gas change for the last ten years?”) and refine his research interactively. This project is at the junction of several distinct research communities (information retrieval, spoken language dialog systems, natural language generation) and is not only about integrating existing tools but also and mainly about studying a newly created, emerging object, a new kind of human-computer dialog. In particular it includes collaborative information search and dynamic co-building of semantics and interaction domain.

One of the first step of this project was to collect a corpus of spoken queries. In this paper we will present the methodology used to collect and to annotate the corpus. This corpus should eventually be available to the community.

We developed a first platform (Galibert et al., 2005) to collect data. The main components of the Ritel system are the speech recognizer, the entities tagger, the dialog manager, the question-answering system, the natural language generation system and the Text-To-Speech synthesizer. They communicate through a message-passing infrastructure. The dialog manager controls and organises the interaction. It manages the entities tagger and the information passed through to the QA system. An overview of the spoken dialog system architecture is shown in Figure 1.

The dialog manager was designed to incite people to talk as much as possible, to reformulate their requests in many ways, and refining their question while keeping a reasonably natural interaction. It can hence be considered an Eliza variation (Weizenbaum, 1966). Moreover it allows searching of information in databases as appropriate. Each semantic frame sent by the specific entity detector is 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 the specific request for information and refinement of the request occurs). If according to these models the current request is considered to be of the kind that can be answered factually by searching in one of the available databases, the search module extracts 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 that the system has understood something and urges the user to refine or reformulate the question. These two modules generate new semantic frames that are sent to the natural language generation (NLG) module. 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.

2. Corpus description
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 whatever they want however they want. Of the 6 hours of user speech one hour has been set aside for development (dev) and one hour for testing (test) purposes.

The total corpus contains 6 hours of user speech, 5360 user queries in 582 dialogs. Table 1 gives an overview of the corpus.

All the corpus has been orthographically transcribed and annotated in terms of topic and specific entity. See for
instance the following utterance:

qui détient le rôle principal dans le grand alibi est -ce un homme ou une femme (who is the main actor in the grand alibi is it a man or a woman)

2.1. Topic annotation

The topic are hierarchically organized. There are 12 main topics (such as animal, arts, music, sciences etc.). These main topics are subdivided into 22 sub-topics (such as vocabulary, biology, law etc.). 536 utterances received a null topic, 4824 one main-topic and 63 two main-topics. In all of the dual-topic queries, the first topic reference is a rejection by the user of a misunderstanding of the system and the second the real object of the query. Additionally, 1171 sub-topics have been annotated.

Table 7 shows examples for each main-topic. Each topic can be refined with one or more sub-topics which have grown from the actual contents of the corpus. Table 2 show the 10 most present full topic classifications found in the corpus.

| Succession                                      | # Occ. |
|------------------------------------------------|--------|
| culture_generale/politique                      | 387    |
| culture_generale/societe/economie               | 134    |
| culture_generale/vie_pratique/vocabulaire      | 83     |
| culture_generale/vie_pratique                   | 77     |
| arts/peinture                                  | 63     |
| science/astronomie                              | 30     |
| culture_generale/societe                        | 21     |
| science/biologie                               | 20     |
| science/physique                               | 13     |
| arts/architecture                               | 13     |

Table 2: 10 most frequent successions of topics and sub-topics

2.2. Named and Extended Entities annotation

The specific entities annotated in the corpus are from 5 categories:

- Standard named entities such as people, products, titles, commercial names, time markers, organizations and places. Table 3 shows examples of the different categories of standard NEs.

- General entities like lexical units, general amount, activity, status, animal, sport, geographical origin, citation and administrative function. Table 4 shows examples of these different categories.

- Extended entities which covers unspecified named entities (such as "the Olympic Games" instead of "the Olympic Games of 1992 in Barcelona"). Examples are shown in Table 5.

- Topic-defining entities such as history, literature, politics, sciences... Examples are shown in Table 8.

- Linguistic entities such as specifiers, superlatives, comparatives. Table 6 shows examples of these categories.

3. Conclusion

This corpus is a real (as opposed to Wizard-of-Oz) Human-Computer QA-oriented spoken dialog corpus. The user utterances have been fully transcribed and annotated and we working towards its free distribution to the community. Some, but by no means all, of the expected uses for it are:

- Dialog System development. The data is usable as-is for both the speech recognition side (acoustic and language models) and the dialog management side (automatic understanding)

- Natural Language Question Answering. A number of user strategies towards interactive information retrieval can be seen through the corpus, including question explanations and reformulations.

- Named Entities Detection.
| Category | Example |
|----------|---------|
| LOC      | quelle est la plus grande ville du Soudan | quel est le plus grand Chirac |
| PERS     | je voudrais des informations sur Fritz Lang | qui détient le rôle principal |
| ORG      | quels pays font partie de l'Europe | Chirac il est de quelle nationalité |
| TIME     | quel nom a porté la ville de Saint-Pétersbourg | la fin des années 60 |
| PROD     | qui a écrit le rouge et noir | qui a écrit le rouge et le noir |
| EVENT    | au festival de Cannes | où vont les prochains jeux olympiques |

Table 3: Examples of the different categories of Named Entities

| Category | Example |
|----------|---------|
| UL       | que veut dire le mot <ul> diaspora </ul> | quelle est la plus grande ville de Soudan |
| CIT      | who said <cit> a good conscience is a continual Christmas </cit> | qui est le personnage principal du livre |
| SPORT    | question de sport question de <sport> natation | quel est le personnage principal du livre |
| ANIMAL   | j’aimerais savoir si une <animal> puce </animal> | j’aimerais savoir si une <animal> puce |
| AMOUNT   | ... fait des bonds de 19 centimètres | ... fait des bonds de 19 centimètres |
| ORIG     | quel roi <orig> anglais </orig> | quel roi <orig> anglais |
| FONCTION | quel <fonction> roi </fonction> | quel <fonction> roi |
| STATUS   | quel est le <status> plus grand </status> | qui est le <status> plus grand |
| ACTIVITY | quel est le plus grand <activity> sculpteur </activity> | qui est le plus grand sculpteur |

Table 4: Examples of the different categories of Extended Entities

| Category | Example |
|----------|---------|
| Loc      | quelle est la plus grande ville | quelle est la plus grande ville du Soudan |
| Prod     | dans quel <Prod> film | dans quel <Prod> film |
| Pers     | qui détient le rôle principal | qui détient le rôle principal |
| Eve      | où ont lieu les prochains jeux olympiques | où ont lieu les prochains jeux olympiques |
| Org      | Chirac il est de quelle nationalité | Chirac il est de quelle nationalité |
| Time     | la fin des années | la fin des années 60 |

Table 5: Examples of the different categories of Imprecises Entities

| Category | Example |
|----------|---------|
| Status   | quelle est la ville du Soudan | quelle est la ville du Soudan |
| Spec     | quel est le personnage principal du livre | quel est le personnage principal du livre |
| Objquest | le film où on voit un homme accroché ... | le film où on voit un homme accroché ...

Table 6: Examples of the different categories of Linguistic Entities

4. References

O. Galibert, G. Illouz, S. Rosset. 2005. Ritel: An Open-Domain,Human-Computer Dialog System. In Proc. of Interspeech’05. 2789-2792.

J. Weizenbaum. 1966 ELIZA: A Computer Program For the Study of Natural Language Communication Between Man and Machine. In Communications of the ACM, 9(1), 36-35.
Table 7: Topics and Examples

| Topic          | Example                                                                 | # Occ. |
|----------------|-------------------------------------------------------------------------|--------|
| Music          | non je m' intéresse aux musiques de fi lm no I’m interested in movie music | 41     |
| History        | je voudrais des informations sur Louis XIV I would like informations about Louis XIV | 314    |
| Geography      | je voudrais savoir la capitale du Venezuela what is the capital of Venezuela | 1637   |
| Science        | comment s’ appelle le gros télescope qui est dans l’ espace what is the name of the large telescope which is in space | 140    |
| Film           | qui a obtenu le dernier oscar du meilleur fi lm who won the last oscar for the best movie | 734    |
| Literature     | je cherche des informations sur Beaudelaire I’m looking for information on Beaudelaire | 411    |
| Sport          | le nombre de joueurs dans une équipe de foot gaélique how many players are in a wales football team | 132    |
| Animal         | sur la reproduction des tortues de mer on the reproduction of the Sea turtles | 57     |
| Arts           | qui a peint l’ Angelus who painted the Angelus | 124    |
| General Culture| une information sur le prix nobel information on nobel prize | 792    |
| Closing        | au-revoir bye | 378    |
| Opening        | allô allo | 1      |

Table 8: Examples of the different categories of Topic-defining Entities