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 (f.i. "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 studiying 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 developped 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 synthetizer. 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.

# dial.	duration of user speech	# user queries	# user 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
1171	4761
# dist. topics	# dist.subtopics
12	22

# Entities	# Distinct entities
10378	63

Table 1: Summary of the RITEL corpus

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 entities. See for

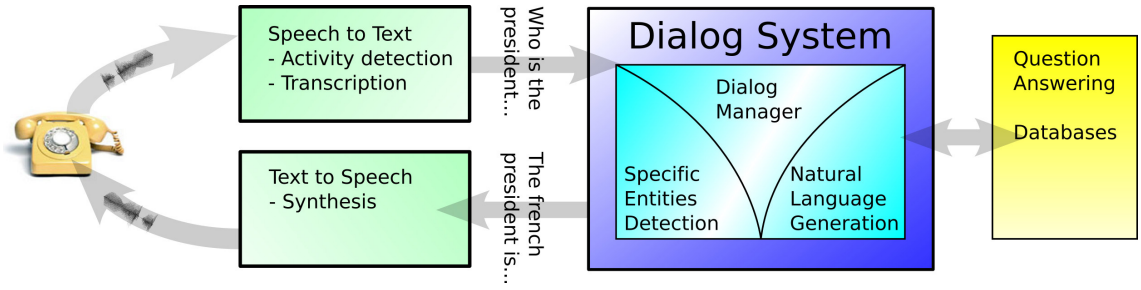

Figure 1: Overview of the RITEL system

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)

$|1\_cinema||$ qui détient le $< Pers >$ rôle $< /Pers >$ $< spec >$ principal $< /spec >$ dans $< prod >$ le grand alibi $< /prod >$ est -ce un $< Pers >$ homme $< /Pers >$ ou une $< Pers >$ femme $< /Pers >$

Where $|1\_cinema|$ is the first main topic
$< Pers >$, $< spec >$ and $< prod >$ are specific entities.

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 entites 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 example 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 <loc> Soudan </loc> what is the biggest town in <loc> Soudan </loc>
PERS	je voudrais des informations sur <pers> Fritz Lang </pers> I'd like informations on <pers> Fritz Lang </pers>
ORG	quels pays font partie de l' <org> Europe </org> what coutries are in <org> Europe </org>
TIME	quel nom a porté la ville de Saint Petersbourg jusqu'en <time> 1991 </time> what was the name of Saint Petersbourg until <time> 1991 </time>
PROD	qui a écrit <prod> le rouge et le noir </prod> who wrote <prod> the red and black </prod>
EVENT	au <eve> festival de Cannes en 1983 </eve> ... in the <eve> Cannes film festival in 1983</eve> ...

Table 3: Examples of the different categories of Named Entities

Category	Example
UL	que veut dire le mot diaspora what is the meaning of the word diaspora
CIT	who said <cit> a good conscience is a continual Christmas </cit>
SPORT	question de sport question de <sport> natation </sport> sports question, about <sport> swimming </sport>
ANIMAL	j' aimerais savoir si une <animal> puce </animal> ... I'd like to know whether a <animal> flea </animal> ...
AMOUNT	... fait des bonds de <val> 19 centimètres </val> can do <val> 19 centimeters </val> jumps
ORIG	quel roi <orig> anglais </orig> ... which <orig> english </orig> king ...
FONCTION	quel <fonction> roi </fonction> ... which <fonction> king </fonction> ...
STATUS	quel est le <status> plus grand </status>... who is the <status> most famous </status>...
ACTIVITY	quel est le plus grand <activity> sculpteur </activity>... who is the most famous <activity> sculptor </activity>...

Table 4: Examples of the different categories of Extended Entities

Category	Example
Loc	quelle est la plus grande <Loc> ville </Loc>... which is the largest <Loc> town </Loc>...
Prod	dans quel <Prod> film </Prod> ... is which <Prod> movie </Prod> ...
Pers	qui détient le <Pers> rôle</Pers> principal who is the main <Pers> actor</Pers>
Eve	où ont lieu les prochains <Eve> jeux olympiques </Eve> where will the next <Eve> olympic games </Eve> be
Org	Chirac il est de <Org> gauche </Org> ou ... Chirac is he <Org> left-wing </Org> or ...
Time	la fin des <Time> années 60 </Time> the end of the <Time> sixties </Time>

Table 5: Examples of the different categories of Imprecises Entities

Category	Example
Status	quelle est la <statut> plus grande </statut> ville du Soudan what is the <statut> largest </statut> town in Soudan
Spec	quel est le personnage <spec> principal </spec> du livre who is the <spec> main </spec> hero in the book
Objquest	le film où on voit un homme <objquest> accroché à une aiguille de pendule </objquest> the movie where you can see a man <objquest> clinging to a clock hand </objquest>

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.

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 téléscope 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 7: Topics and Examples

Category	Example
Tnom	je cherche le <Tnom> nom </Tnom> d' un peintre I'm looking for the <Tnom> name </Tnom> of a painter
Ttime	à quel <Ttime> époque </Ttime> a été construit... in what <Ttime> period </Ttime> was it built...
Tmesure	quel est l' <Tmesure> âge </Tmesure> de Robert Redford <Tmesure> how old </Tmesure> is Robert Redford
Tpopulation	<Tpopulation> combien y a d' habitants </Tpopulation> à Libreville <Tpopulation> how many inhabitants </Tpopulation> in Libreville
Tdatenaiss	et quand il est <Tdatenaiss> né </Tdatenaiss> and when was he <Tdatenaiss> born </Tdatenaiss>
Tdatemort	et la <Tdatemort> date de son décès </Tdatemort> and what is the <Tdatemort> date of his death </Tdatemort>
Tsuperficie	je voudrais connaître la <Tsuperficie> superficie <Tsuperficie> en kilomètres carrés de la France I'd like to know the <Tsuperficie> size <Tsuperficie> in square kilometers of France
Tlangue	les <Tlangue> langues officielles </Tlangue> en Europe the <Tlangue> official languages </Tlangue> in Europe
Tmonnaie	quelle est la <Tmonnaie> monnaie </Tmonnaie> utilisée en Tunisie what <Tmonnaie> currency </Tmonnaie> is used in Tunisia
Torthographe	la <Torthographe> graphie </Torthographe> de timbre-poste <Torthographe> how do you write </Torthographe> post-stamp
Tvocety	je voudrais savoir quelle est son <Tvocety> étymologie </Tvocety> I'd like to know its <Tvocety> ytymology </Tvocety>
Tvocens	<Tvocsens> quel est le sens</Tvocens> du mot diaspora <Tvocsens> what is the meaning </Tvocens> of the word diaspora
Tnationalite	quelle est la <Tnationalite> nationalité </Tnationalite> de cet acteur what is the <Tnationalite> nationality </Tnationalite> of that actor
Tprofession	quelle était la <Tprofession> specialité </Tspecialite> de Phidias what was the <Tprofession> speciality </Tspecialite> of Phidias
Tclimat	quel est le <Tclimat> climat </Tclimat> au Brésil what is the <Tclimat> weather </Tclimat> like in Brasil

Table 8: Examples of the different categories of Topic-defi ning Entities