The French MEDIA/EVALDA project: the evaluation of the understanding capability of Spoken Language Dialogue Systems

L. Devillers (1), H. Maynard (1), S. Rosset (1), P. Paroubek (1), K. McTait (2), D. Mostefa (2), K. Choukri (2), L. Charnay (3), C. Bousquet (4), N. Vigouroux (4), F. Béchet (5), L. Romary (6), J.Y Antoine (7), J. Villaneau (7), M. Vergnes (8), J. Goulian (9)

LIMSI-CNRS (1), ELDA/ELRA (2), FRANCE-TELECOM R&D (3), IRIT (4), LIA (5), LORIA (6), VALORIA (7), VECSYS (8), CLIPS (9)

media@elda.fr

Abstract

This paper presents and reports on the progress of the EVALDA/MEDIA project, focusing on the recording and annotating protocol of the reference dialogue corpus. The aim of this project is to design and test an evaluation methodology to compare and diagnose the context-dependent and independent understanding capability of spoken language dialogue systems. Systems from both academic organisations (IRIT, LIA, LIMSI, LORIA, VALORIA, CLIPS) and industrial sites (FRANCE TELECOM R&D, TELIP) will be tested. ELDA is the coordinator of the Technolangue/EVALDA multi-campaign evaluation project, of which MEDIA is a sub-campaign. LIMSI is the scientific coordinator of the project. MEDIA began in January 2003.

Introduction

The aim of the MEDIA project is to design and test a methodology for the evaluation of context-dependent and independent spoken dialogue systems. We propose an evaluation paradigm based on the use of test suites from real-world corpora and a common semantic representation and common metrics. This paradigm should allow us to diagnose the context-sensitive understanding capability of dialogue systems. This paradigm will be used within an evaluation campaign involving several sites all of which will carry out the task of querying information from a database.

Presently, there are no common standard methodologies or practices agreed upon by the scientific community for the evaluation of spoken dialogue systems. The dynamic and interactive nature of dialogue makes it difficult to construct a reference corpus of dialogues against which systems may be evaluated. On the other hand, various influential projects have tried to build the foundations of en evaluation methodology for spoken dialogue systems, beginning with the francophone project AUF-Arc B2 (Mariani 1998), the evaluation carried out by DEFI (Antoine 2002), the European EAGLES projects (Dybkjaer 1998) and subsequently DISC (Giachim 1997), SUNDIAL (Gibbon 1997), and the ATIS (MADCOW 1992) and COMMUNICATOR (Walker 2001) projects in the USA.

In the PEACE paradigm (Paradigme d'Evaluation Automatique de la ComprEhension hors- et en- contexte dialogique) (Devillers 2002, Maynard 2000) on which the MEDIA project is based, it is possible to carry out an automatic, comparative and diagnostic evaluation of the context-dependent and independent understanding capability of a dialogue system. It is based on the construction of reproducible test suites from real dialogues. This paradigm follows the same idea as the DQR (Antoine 2000) and DEFI (Antoine 2002) evaluations based on test suites. The evaluation environment relies on the premise that, for database query systems, it is possible to construct a common semantic representation to which each system is capable of converting its own internal representation. Within this paradigm, it is also possible to carry out a context-dependent evaluation. The context is artificially simulated, by paraphrasing, with the aim of testing an utterance U in the context D^n (using DQR notation). Finally, while the large evaluation programmes centred on performance evaluation (global measures), this campaign will not only make possible performance but also a diagnostic evaluation of the models used. Therefore, the objective of the MEDIA project is to give the francophone scientific community the means with which they can make comparative evaluations of understanding modules while offering them the possibility to share corpora and define representations and generic common metrics.

The first stage of the MEDIA project has been dedicated to defining, constructing and then annotating a common corpus of dialogues in French relevant to the task chosen for the MEDIA project (tourist information server). After a presentation of the MEDIA project and of the state of our common semantic annotation scheme, this paper will present the methodology used for constructing the corpus (task definition, description of the recording platform, recording, annotating protocol, etc).

The MEDIA project

2.1 Organisation of the campaign

The aim of organising an evaluation campaign to test the understanding capability of context-dependent and independent dialogue systems is to promote an evaluation framework for the scientific community. The aim of this project is to establish a generic evaluation paradigm to test the context-dependent and independent understanding capability enabling an automatic, comparative and diagnostic system evaluation. An evaluation campaign should ensure that the resources created as well as all side products of the project are of a lasting and permanent nature. To ensure impartiality, the campaign is coordinated and managed by ELDA who is not participating in the evaluation campaign. ELDA is also in charge of creating the corpus necessary for the project and is responsible for creating or providing the software or tools necessary for the evaluation campaign itself. The company VECSYS has provided the recording platform for the corpus (hardware and software including the WoZ system). The initiator of the project, LIMSI, is responsible for coordinating the scientific aspects of the project. Participants from both academic (IRIT, LIA, LIMSI, LORIA, VALORIA, CLIPS) and industrial sites (France Telecom R&D, TELIP) will take part in the system evaluations.

2.2 Evaluation Paradigm

In order to provide a diagnostic evaluation, the evaluation paradigm relies on a common generic semantic representation. The formalism chosen will be agreed upon by all project partners and must enable a large corpus to be decorated with both context-dependent and context-free annotations. The results of our first discussions on the domain-independent common annotation scheme are presented in this paper.

A common generic semantic representation

This involves establishing a representation of the meaning of the user utterances enabling the relationship between corresponding equivalences in the set of possible requests to be visualised. However, it is to be defined with respect to the database query task. The common semantic representation is based on a attribute-value structure in which conceptual relationships are implicitly represented by the name of the attributes. This formalism enables communicative acts as well as the semantic content of an utterance to be coded in a two levels attribute-value representation. The communicative acts are derived from FIPA (FIPA 2002) Communicative Act Library (CAL). Six dialog acts have been agreed by all participants: Inform, Query, Accept (Confirm), Reject (Disconfirm), Opening and Close. This reduced list allows obtaining an high level of inter-annotators agreement. The proposed semantic representation, based on (Maynard 2003), consists of a hierarchy of attributes, which are identified in an attribute dictionary. This conceptual hierarchy provides also a set of relationships between semantic units. Each turn (client and agent) of a dialog is segmented into one or more dialogic segments and each dialogic segment is segmented into one or more semantic segment with the assumption that a semantic segment corresponds to a single attribute.

A semantic segment is represented by a 5-tuplets which contains: the mode (positive, affirmative, interrogative or optional), the name of the attribute corresponding to the segment, the value of the attribute, links: optional pointers to related segments in the dialog and an optional comment on the segment. The order of the 5-tuplets in the semantic representation follows their order in the utterance. The values of the attributes are either numeric units, proper names or semantic classes merging lexical units, which are synonyms for the task. This Attribute-Value Representation (AVR) allows simple annotation process. It is planned to build a Feature Structure from this AVR in order to represent explicitly the relationships between segments.

Reference units

A unit of reference for the evaluation of context-independent understanding capabilities consists of the exact orthographic transcription of the user utterances and the reference semantic representation. A unit of reference for the evaluation of context-dependent dialogue understanding capabilities consists of the context in the form of a paraphrase (Devillers 2002), the exact orthographic transcription of the user utterance and the semantic representation resulting from the interpretation of the utterance taking the context into account. The paraphrase may be obtained by concatenating the user utterances and, optionally by including the system responses. The set of units of reference will be divided into 3 sections: a training corpus (10k requests), a development corpus (2k requests) distributed to participants and an unseen test corpus (3k requests) for the evaluation itself. Each user utterance transcribed according to the usual transcription norms for oral utterances is annotated according to the context-dependent and independent common semantic representation.

Common evaluation metrics

The aim is to define common metrics in order to be able to carry out a diagnostic system evaluation. It must also be possible to balance the importance of the errors according to dialogic phenomena.

Definition and typology of the dialogic and linguistic phenomena

The paradigm must offer a qualitative analysis and automatic diagnosis of the performance of the context-dependent and independent understanding module. For example, it would be possible to study the particular difficulties associated with speech, independent of the context i.e. hesitation, repetition etc.

Corpus construction

3.1 Task and domain definition

Within an evaluation campaign for man-machine dialogues, it is necessary to restrict the database query

task to querying a tourist information server, train/aeroplane timetable information server etc. The definition of the semantic representation is generic. It is then adapted to suit the task and the database. The ideal scenario would be to work on an application connected to real-world database, for example, an interface to the website of a travel agent or tourist office. The common task chosen for this evaluation is the reservation of a hotel room with tourist information using information obtained from a web-based database.

3.2 Data collection

It is necessary to have a corpus of common dialogues in order to train the different context-dependent dialogue systems and to create the test suites for the evaluation. In order to avoid bias in the evaluation, we decided to record a new corpus simulating a vocal tourist information server by a Wizard of Oz (WoZ) system. In this way, each user/caller believes he or she is talking to a machine whereas in actual fact he is talking to a human being (a 'wizard') who simulates the responsorial behaviour of tourist information server. This enables us to obtain a corpus of varied dialogues due in part to the behaviour of the wizard.

In this campaign, it was decided that only the orthographic transcriptions of the speakers (caller and system/wizard) would be used as the basis for the evaluation.

However, it would also be advantageous to have the high quality (digital) audio signal available with the aim of expanding the campaign, at a later date, to include speech recognition systems capable of providing the input to such dialogue systems. The proposed size of the corpus is approximately 15,000 user requests. To achieve this figure, 1250 dialogues are to be recorded, using 250 different callers where each caller carries out 5 different reservation scenarios. The final corpus will be of the order of 70h of dialogue.

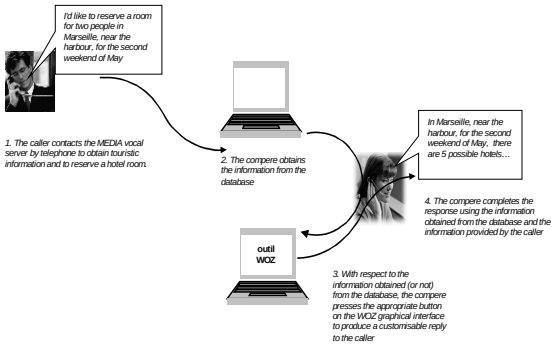

Figure 1 : *The WoZ System*

Recording Platform

The method chosen for the corpus construction process is that of a 'Wizard of Oz' (WoZ) system. The operator (wizard) uses the graphical interface, developed by VECSYS, which assists him to generate responses that are to be communicated to the caller. Figure 1 illustrates the architecture of the recording platform.

The generated replies are obtained by completing a sentence template with the information obtained by consulting a tourist information website taking into account the speaker's request. The signal is recorded in digital format. The dialogues are subsequently transcribed orthographically then separated into dialogic acts and annotated semantically.

Recording protocol

The callers that take part in the recording refer to pre-defined tourist and hotel reservation scenarios, generated from a set of templates in such a way as to have a set of varied dialogues. In order to obtain a set of user requests made in as natural way as possible, the scenarios are given to the callers by telephone. This reduces the amount of paraphrase or repetition from the scenarios in textual form. Several starting points are possible for the dialogues i.e. choice of town, itinerary, touristic event, festival, price, date etc. Eight scenario categories were defined each with a different level of complexity. An example of a simple scenario (translated from French) is given in Figure 2. A complex scenario could consist of reserving several hotels in several locations according to an itinerary.

DATE:	2nd weekend of May
TOWN:	Marseille
SITUATION:	Near the harbour
No.ROOMS :	1 single room
No.ADULTS :	1
PRICE:	50-60 Euros per night

Figure 2 : *A simple scenario*

In addition to the variety of scenarios given to the callers, we defined a set of instructions for the wizard in order to respond to the caller/user requests. There are three categories of instructions. The first concerns speech recognition or comprehension errors. In this way, the wizard produces a response having 'misunderstood' the user request. The second involves explicit or implicit feedback to the user. The final type concerns the level of cooperation on the part of the wizard. On one end of the spectrum, the wizard returns all the information requested by the user. On the other end he is not able to reply to any of the user's requests. Between these two extremes, the wizard may provide partial information to the user. In addition to the instructions given to the wizard, the user is given instructions as to the number and type of parameters he or she can negotiate with the server.

3.3 Current state of corpus

The corpus is currently being recorded. At the present time, approximately 4/5 of the corpus (1000 dialogs) has been produced by two wizards. Table 1 indicates the average measurements of the corpus transcribed (200 dialogs) so far in terms of the average number of user utterances in a dialogue, average duration of a dialogue, length of a user utterance and length of a system utterance. The size of the user lexicon is around 1.1 K words

No. dialogs transcribed	200 dialogs
Average length	3.4 minutes
No. user utterances per dialogue	15 utterance
Length user utterance	6 words
Length system utterance	10 words

Table 1: *Preliminary observations on the corpus*

The diversity of the utterances produced depends not only on the complexity of the scenario, but also on the behaviour of the wizard. The most interesting phenomena (reference, negotiation, negation) are observed above all during complex scenarios with a non-cooperative Wizard. Figure 3 presents an extract of a dialogue (translated from French). From this extract, one can observe phenomena frequently found in a dialogue such as hesitation, repetition, as well as references such as 'that night' or 'the same thing'.

```
U : err, well, ah, I'd like to reserve
a room for that night, so two rooms at
the Mercure Hotel err in Lille.

S: Reserving two rooms at the Mercure
Hotel, the Grand Hotel in Lille. Do you
require any further information?

U: well, that'll be the same thing so
two rooms also in Paris err for the
nights of the 21st and also the 22nd of
February with the same err with the
same criteria err so still two couples
with one child
```

Figure 3 : *Example dialogue*

Corpus distribution

The corpus, including the transcriptions and the semantic annotations, will be distributed by ELRA/ELDA as widely as possible in the form of an *evaluation package*, which will also contain the (anonymous) evaluation results and the tools developed for the campaign. The consortium will pay attention to the concept of reusability of resources with the aim of contributing to the standardisation of testing methods. The aim of this distribution is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

Conclusion

The data collection phase is due to finish by the end of March 2004. The transcription of these dialogues will be finished shortly after. Each dialogue (audio signal and transcriptions) will be accompanied with the set of instructions given to the caller and the wizard. Currently, work is taking place on analysing the dialogues already recorded with the aim of finalising the structure of the common semantic representation and the set of task-dependent concepts. The semantic annotation of the dialogues will start in April 2004.

Acknowledgements

The MEDIA evaluation campaign forms part of the French language engineering evaluation project EVALDA, financed under the French cross-ministerial *Technolangue* initiative.

References

Mariani J. (1998) "The Aupelf-Uref Evaluation-Based Language Engineering Action and Related Projects", *LREC 1998*, vol. 1, Granada. Spain..

Antoine J., & all, (2002) "Predictive and objective evaluation of speech understanding : the challenge evaluation campaign of the I3 speech workgroup of the French CNRS", *LREC 2002*, Las Palmas, Spain.

Dybkjaer L. & all, (1998), "The Disc Approach to Spoken Language System Development and Evaluation", *LREC 1998, Granada, May 1998, vol. 1 pp 185-189.*

Giachin, E. and McGlashan, S. (1997). "Spoken Language Dialogue Systems". In S. Young and G.Bloothooft (Eds.) Corpus-based methods in language and speech processing. Dordrecht: Kluwer Academic Publishers, 69-117.

Gibbon D. Moore R. W. R., (1997), "*Handbook of Standards and Resources for Spoken Language Resources"*, Mouton de Gruyter, New-York, *1997*, ISBN 3-11-015366-1.

MADCOW, (1992) "Multi-Site Data Collection for a Spoken Language Corpus", *DARPA Speech and Natural Language Workshop*, 1992

Walker M., Passonneau R., Boland J. (2001), "Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialog Systems", *ACL/EACL Toulouse, 2001*

Devillers, H. Maynard, P. Paroubek, (2002), "Méthodologies d'évaluation des systèmes de dialogue parlé : réflexions et expériences autour de la compréhension", *TAL 2002*

Maynard H., Devillers L. (2000) L., "A framework for evaluating contextual understanding", *ICSLP* 2000.

Antoine J., & all, (2000) "Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm", *LREC* 2000.

Maynard, H. Rosset. S, (2003) "A Semantic representation for spoken dialogs", *EUROSPEECH 2003*

FIPA Communicative Act Library Specification, SC00037J, FIPA TC Communication, December 3rd 2002, http://www.fipa.org/specs/fipa00037