

Robust spoken dialogue comprehension for human-robot interaction

Pierre Lison

Cognitive Systems @ Language Technology Lab
German Research Centre for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany.

[pierrel@coli.uni-sb.de]

*Studentische Tagung der Sprachwissenschaft (StuTS)
Amsterdam, the Netherlands*

May 2008



Short Introduction

- What I am going to present in this talk is part of the research work for my **M.Sc. thesis in Computational Linguistics** (at the *Universität des Saarlandes*, in Saarbrücken).
- Its (current) title is :
 - « *Robust spoken dialogue comprehension for human-robot interaction* ».
- That is, I seek to develop techniques & algorithms to allow a robot to understand spoken dialogue in a **robust** fashion.
- The system is currently being implemented as part of a cognitive architecture for mobile robots interacting with humans to perform a variety of service-oriented tasks.



Preliminary notes

- It is still a **work in progress** !
- The system is only partially implemented, and for the time being I only have experimental results for some subcomponents.
- *Your comments, suggestions, criticisms are very welcome !*
- **Acknowledgements** : Part of the material of this talk is drawn from [Kruijff 06] and [Kruijff 07].



Outline of the talk

- 1 Background
 - Human-robot interaction (HRI)
 - Cognitive systems for HRI
 - Spoken dialogue comprehension
- 2 Approach
 - Part 1 : Context-sensitive speech recognition
 - Part 2 : Robust parsing
- 3 Conclusions
- 4 Bibliography



Outline of the talk

1 Background

- Human-robot interaction (HRI)
- Cognitive systems for HRI
- Spoken dialogue comprehension

2 Approach

- Part 1 : Context-sensitive speech recognition
- Part 2 : Robust parsing

3 Conclusions

4 Bibliography



Outline of the talk

1 Background

- Human-robot interaction (HRI)
- Cognitive systems for HRI
- Spoken dialogue comprehension

2 Approach

- Part 1 : Context-sensitive speech recognition
- Part 2 : Robust parsing

3 Conclusions

4 Bibliography



Outline of the talk

1 Background

- Human-robot interaction (HRI)
- Cognitive systems for HRI
- Spoken dialogue comprehension

2 Approach

- Part 1 : Context-sensitive speech recognition
- Part 2 : Robust parsing

3 Conclusions

4 Bibliography



Talking robots ?

- Our long-term aim :

« *Hi, I am C3-PO, Human Cyborg Relations.* »



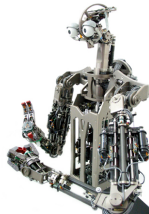
(And he knows over 6 million languages...)

- For the time being, we'll obviously need to scale down our expectations...



Today's "state of the art"

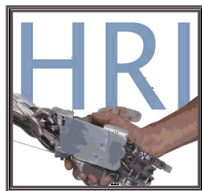
- Nevertheless, research in robotics is rapidly moving forward, and we are already able to do a few things :





Human-Robot Interaction

- How to make robots that actually *understand* what we say? And that understand *why*, *when* and *how* they should say something?
- Research in HRI seek to develop principles and techniques to allow *efficient* and *natural communication* between robots and humans
- *Interdisciplinary* research field : artificial intelligence, robotics, (computational) linguistics, & the social sciences – psychology, cognitive science, anthropology, etc.





Human-Robot Interaction (cont'd)

- HRI is always about *situated* interaction : Language often refers to reality, discusses actions & plans affecting that reality.
- **Situated dialogue understanding** is thus crucial for HRI.
= Understanding and producing language, relative to a current or imaginable situation in which the agents are situated.
- It means that we *cannot* consider communication in isolation from the other modalities – we need to find meaningful ways to relate **language**, **action** and **situated reality**.

⇒ need to develop artificial **cognitive systems** able to integrate all these aspects into a common architecture.



Cognitive systems

What is cognition ?

- Cognition is more than intelligence...
- ... it is intelligence set in *reality*.
- “Cognition = perception + intelligence”

What is a cognitive system ?

- A cognitive system is a (artificial or biological) system able to actively *perceive* the environment it find itself in, *reason* about it and *achieve goals* through *plans* and *actions*.



Embodiment

- *Embodiment* modulates how a system sees, experiences, reality.
- “Cognition = embodiment[perception+intelligence]”
- Since they have very different “bodies” (perceptors and actuators), robots and human beings will experience and represent reality in very different ways.
- This difference of embodiment has profound implications for HRI : how can we “create a bridge” between two systems with wildly different conceptions of external reality ?



Cognitive architectures

- Software architectures for cognitive robots are typically composed of several *distributed* and *cooperating* subsystems, such as communication, computer vision, navigation and manipulation skills, and various deliberative processes like symbolic planners.
- All these subsystems are highly interdependent.
- The architecture must moreover enable the robot to use its rich perceptual experience to continuously *learn* and *adapt itself* to the environment.



The CoSy architecture

- Our approach has been implemented as part of a *distributed cognitive architecture*. [Hawes 07].
- Each subsystem consists of a number of processes, and a working memory.
- The processes can access sensors, effectors, and the working memory to share information within the subsystem.
- In this talk we will focus on the subsystem for spoken dialogue comprehension.



Levels of spoken dialogue comprehension

Different levels of processing :

- **Auditory** : speech recognition, (speaker localization & tracking)
- **Grammatical** : syntactic structure, semantic structure
“A grammar specifies the relation between well-formed syntactic structures and their underlying (linguistic) meaning”
- **Discourse** : contextual reference resolution (anaphora, ellipsis), rhetorical relation resolution, (clarification triggers)
“Discourse interprets utterance meaning relative to the established context, establishing how it contributes to furthering the discourse”



Open challenges

- **Robustness** in speech recognition :
 - noise, speaker independence, out-of-vocabulary words
 - poor performance of current ASR technology
 - (intonation, emotion)
- **Robustness** to ill-formed utterances :
 - partial, ungrammatical or extra-grammatical utterances
 - presence of various disfluencies (filled pauses, speech repairs, corrections, repetitions, etc.) in spoken dialogue.
- Pervasive **ambiguity** at all processing levels (lexical, syntactic, semantic, pragmatic)
- **Uncertainty** in contextual interpretation of utterances



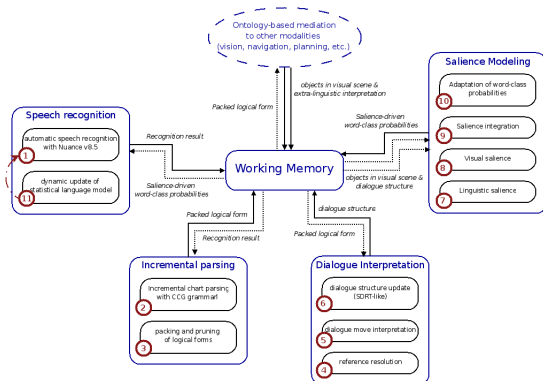
Disfluencies in spoken dialogue : example

- Some interesting examples taken from a corpus of task-oriented spoken dialogue :
The Appolo Lunar Surface Journal.



Architecture for spoken dialogue comprehension

- Schematic view of the CoSy architecture we are currently developing for spoken dialogue comprehension :





Spoken dialogue comprehension : steps

- 1 The speech recognition utilises a commercial system together with a statistical language model, dynamically updated at runtime to adapt itself to the environment.
- 2 Syntactic parsing is based on an incremental chart parser for Combinatory Categorical Grammar, and yields a set of alternative interpretations expressed as ontologically rich, relational structures.
- 3 These are then *packed* into a single representation, a technique which enables us to efficiently handle syntactic ambiguity.



Spoken dialogue comprehension : steps

- 1 The speech recognition utilises a commercial system together with a statistical language model, dynamically updated at runtime to adapt itself to the environment.
- 2 Syntactic parsing is based on an incremental chart parser for Combinatory Categorical Grammar, and yields a set of alternative interpretations expressed as ontologically rich, relational structures.
- 3 These are then *packed* into a single representation, a technique which enables us to efficiently handle syntactic ambiguity.



Spoken dialogue comprehension : steps

- 1 The speech recognition utilises a commercial system together with a statistical language model, dynamically updated at runtime to adapt itself to the environment.
- 2 Syntactic parsing is based on an incremental chart parser for Combinatory Categorical Grammar, and yields a set of alternative interpretations expressed as ontologically rich, relational structures.
- 3 These are then *packed* into a single representation, a technique which enables us to efficiently handle syntactic ambiguity.



Spoken dialogue comprehension : steps

- 4 Once the packed logical form is built, it is retrieved by the dialogue recognition module, which performs dialogue-level analysis tasks such as discourse reference resolution and dialogue move interpretation.
- 5 Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the “binder”, which is responsible for the ontology-based *mediation* across modalities



Spoken dialogue comprehension : steps

- 4 Once the packed logical form is built, it is retrieved by the dialogue recognition module, which performs dialogue-level analysis tasks such as discourse reference resolution and dialogue move interpretation.
- 5 Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the “binder”, which is responsible for the ontology-based *mediation* across modalities



Short recap'

- What have we seen so far?
- In the previous section, we explained :
 - ① The general principles of *human-robot interaction* ;
 - ② Why the development of *integrated cognitive systems* is crucial in enabling robots to interact naturally in situated dialogues ;
 - ③ What is *spoken dialogue comprehension*, why it is so difficult to achieve in real-world environments, and how the processing levels are currently implemented within the CoSy architecture.
- I'll now present my own (modest) contribution to the current research in the field.



Short recap'

- What have we seen so far?
- In the previous section, we explained :
 - ① The general principles of *human-robot interaction* ;
 - ② Why the development of *integrated cognitive systems* is crucial in enabling robots to interact naturally in situated dialogues ;
 - ③ What is *spoken dialogue comprehension*, why it is so difficult to achieve in real-world environments, and how the processing levels are currently implemented within the CoSy architecture.
- I'll now present my own (modest) contribution to the current research in the field.



Short recap'

- What have we seen so far?
- In the previous section, we explained :
 - ① The general principles of *human-robot interaction* ;
 - ② Why the development of *integrated cognitive systems* is crucial in enabling robots to interact naturally in situated dialogues ;
 - ③ What is *spoken dialogue comprehension*, why it is so difficult to achieve in real-world environments, and how the processing levels are currently implemented within the CoSy architecture.
- I'll now present my own (modest) contribution to the current research in the field.



Short recap'

- What have we seen so far ?
- In the previous section, we explained :
 - ① The general principles of *human-robot interaction* ;
 - ② Why the development of *integrated cognitive systems* is crucial in enabling robots to interact naturally in situated dialogues ;
 - ③ What is *spoken dialogue comprehension*, why it is so difficult to achieve in real-world environments, and how the processing levels are currently implemented within the CoSy architecture.
- I'll now present my own (modest) contribution to the current research in the field.



Short recap'

- What have we seen so far ?
- In the previous section, we explained :
 - ① The general principles of *human-robot interaction* ;
 - ② Why the development of *integrated cognitive systems* is crucial in enabling robots to interact naturally in situated dialogues ;
 - ③ What is *spoken dialogue comprehension*, why it is so difficult to achieve in real-world environments, and how the processing levels are currently implemented within the CoSy architecture.
- I'll now present my own (modest) contribution to the current research in the field.



Original contribution, in two parts : I

- 1 We first present an implemented model for **speech recognition** which relies on *contextual information* about salient entities to prime utterance recognition.

The hypothesis underlying our approach is that, in situated HRI, speech recognition performance can be significantly improved by exploiting knowledge about the immediate physical environment and the dialogue history.

This approach has already been implemented, as part of the CoSy architecture. Evaluation results on a test suite show a statistically significant improvement in recognition rate.



Original contribution, in two parts : II

- 2 Secondly, we present a work in progress exploring the application of *grammar relaxation techniques* to overcome the various parsing problems arising from speech disfluencies, ill-formed utterances, and errors in speech recognition.

We argue for a generic and linguistically sound approach based on a weighted Combinatory Categorical Grammar [CCG] in which the grammatical constraints are **relaxed** by means of non-standard combinators.

A statistical model is applied to select the most likely parse among those licensed by this relaxed grammar. The approach therefore combines both *linguistic* and *statistical* sources of information to yield the most likely utterance interpretation.



The issue

- The first step in comprehending spoken dialogue is *automatic speech recognition* [ASR].
- For robots operating in real-world noisy environments, and dealing with utterances pertaining to complex, open-ended domains, this step is particularly error-prone.
- In spite of continuous technological advances, the performance of ASR remains for most tasks at least an order of magnitude worse than that of human listeners [Lippmann 97]



The idea

- The **intuition** underlying our approach : use information about salient entities in the situated environment and in the dialogue state to prime utterance recognition.
- Our **claim** : in HRI, the speech recognition performance can be significantly enhanced by exploiting knowledge about the immediate physical environment and the dialogue state.



(ex-cursus : psycholinguistic motivation)

- Psycholinguistic studies have shown that humans do not process linguistic utterances in isolation from other modalities.
- Eye-tracking experiments notably highlighted that, during utterance comprehension, humans combine, in a closely time-locked fashion, linguistic information with scene understanding and world knowledge [Knoeferle 06].
- These observations, among others, provide evidence for the *embodied* and *situated* nature of language and cognition [Lakoff 87, Barsalou 99].



Approach

- Practically, we use two main sources of information :
 - ① objects in the perceived *visual scene* ;
 - ② linguistic expressions in the *dialogue history*.
- These objects are then ranked according to their **salience**, and integrated into a **cross-modal salience model**.
- This salience model is then applied to dynamically compute **lexical activations**, which are incorporated into the language model of the speech recogniser.



Lexical activation

- To this end, we define in the robot's knowledge base a *lexical activation network*, which lists, for each possible salient entity, the set of words activated by it.
- The network specifies the words which are likely to be heard when the given entity is present in the environment or in the dialogue history.
- It can therefore include words related to the object denomination, subparts, common properties or affordances.
- The salient entity **laptop** will activate words like 'laptop', 'notebook', 'screen', 'opened', 'ibm', 'switch on/off', 'close', etc.



A simple example

- 1 Let's imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).
- 2 The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a "red ball".
- 3 In the robot's knowledge base, the "red ball" object is associated to words like "ball" like "round", "pick up", etc.
- 4 As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.





A simple example

- 1 Let's imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).
- 2 The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a "red ball".
- 3 In the robot's knowledge base, the "red ball" object is associated to words like "ball" like "round", "pick up", etc.
- 4 As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.





A simple example

- 1 Let's imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).
- 2 The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a "red ball".
- 3 In the robot's knowledge base, the "red ball" object is associated to words like "ball" like "round", "pick up", etc.
- 4 As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.





A simple example

- 1 Let's imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).
- 2 The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a "red ball".
- 3 In the robot's knowledge base, the "red ball" object is associated to words like "ball" like "round", "pick up", etc.
- 4 As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.





Evaluation

- We evaluated our approach using a test suite of 250 spoken utterances recorded during Wizard of Oz experiments.
- The participants were asked to interact with the robot while looking at a specific visual scene.
- We designed 10 different visual scenes by systematic variation of the nature, number and spatial configuration of the objects presented. The interactions could include descriptions, questions and commands.





Evaluation results

Word Error Rate [WER]	Classical LM	Saliency-driven LM
<i>with vocabulary \simeq 300 words</i>	25.48 %	24.74 %
<i>with vocabulary \simeq 500 words</i>	31.74 %	27.87 %

TAB.: Comparative results of recognition performance



The issue

- Spoken dialogue contains numerous speech disfluencies – pauses, speech repairs, repetitions, corrections, etc. —, as well as non-grammatical and extra-grammatical utterances
- Moreover, today's speech recognition technology is particularly error-prone, resulting in missing or wrongly recognized words.
- ... this means we may encounter such kind of utterances :
« err robot now could you pick [up] the b/ the ball or [on] the table no not that one the red one yes that's it »
- **The issue** : can we correctly derive (partial) semantic interpretations from this noisy input, using the grammar at our disposal ?



The idea

- The **intuition** behind our approach : “relax” the grammatical constraints to allow for linguistic phenomena which are not *stricto sensu* grammatical, but are nevertheless frequently found in spoken dialogue.
- Examples :
 - ① insertion of (filled or voiceless) pauses in the text string ;
 - ② insertion of various kinds of discourse markers ;
 - ③ repetition of words ;
 - ④ corrections ;
 - ⑤ combination of several dialogue acts within the same utterance ;

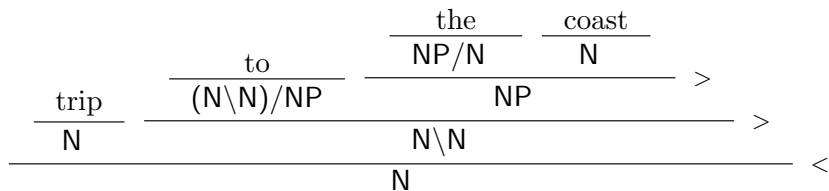


Combinatory Categorical Grammar

- To explain how this grammar relaxation works in practice, I'll first (briefly) the grammar formalism we are using.
- Our grammar is specified as a **Combinatory Categorical Grammar** :
 - Mildly context-sensitive grammar formalism
 - Fully lexicalized
 - Building structures using combinatory rules
 - Very efficient (polynomial parsability) and fully implemented (open source development platform : OpenCCG)



Simple parsing example





Grammar relaxation in CCG

- The step-by-step construction of the parses is triggered using combinatory rules
- For instance, the example we just saw uses the following functional application rules :

$$A/B \ B \Rightarrow A \quad (>)$$

$$B \setminus A \ B \Rightarrow A \quad (<)$$

- The idea is to introduce in the grammar a set of **non-standard combinators** in order to account for the disfluencies and ill-formed utterances.



Constraining the set of parses

- Now we are left with another problem : by inserting additional non-standard combinators into the grammar, we also multiply the number of parses for each utterance.
- And we want to avoid at all cost a *combinatorial explosion* of the number of parses !
- **Possible solution** : Use a *statistical model* to select the most likely parse among those licensed by the grammar.
- The parameters of this statistical model would be estimated on a corpus.
- This is precisely what I'm trying to currently achieve for my M.Sc thesis – first results expected in one or two months :-)



Conclusions

- In this talk, I explained how we could develop robots endowed with *communicative abilities*, ie. artificial agents able to understand situated dialogue.
- We investigated some of the challenges encountered in this task, esp. the lack of *robustness* of many dialogue systems.
- I finally described *two new techniques* I'm currently experimenting to partially alleviate these problems :
 - ① taking the situated context into account to improve the speech recognition performance ;
 - ② relaxing the grammatical constraints to enable the parser to recover (partial) semantic interpretations from ill-formed input.



Conclusions

- In this talk, I explained how we could develop robots endowed with *communicative abilities*, ie. artificial agents able to understand situated dialogue.
- We investigated some of the challenges encountered in this task, esp. the lack of *robustness* of many dialogue systems.
- I finally described *two new techniques* I'm currently experimenting to partially alleviate these problems :
 - ① taking the situated context into account to improve the speech recognition performance ;
 - ② relaxing the grammatical constraints to enable the parser to recover (partial) semantic interpretations from ill-formed input.



Conclusions

- In this talk, I explained how we could develop robots endowed with *communicative abilities*, ie. artificial agents able to understand situated dialogue.
- We investigated some of the challenges encountered in this task, esp. the lack of *robustness* of many dialogue systems.
- I finally described *two new techniques* I'm currently experimenting to partially alleviate these problems :
 - ① taking the situated context into account to improve the speech recognition performance ;
 - ② relaxing the grammatical constraints to enable the parser to recover (partial) semantic interpretations from ill-formed input.



Conclusions

- In this talk, I explained how we could develop robots endowed with *communicative abilities*, ie. artificial agents able to understand situated dialogue.
- We investigated some of the challenges encountered in this task, esp. the lack of *robustness* of many dialogue systems.
- I finally described *two new techniques* I'm currently experimenting to partially alleviate these problems :
 - ① taking the situated context into account to improve the speech recognition performance ;
 - ② relaxing the grammatical constraints to enable the parser to recover (partial) semantic interpretations from ill-formed input.



The end

Thank you for your attention !!





⇒ **Questions, comments ?**

For more information, visit
<http://www.dfki.de/cosy>






Bibliography I

-  L. W. Barsalou.
Perceptual symbol systems.
Behavioral & Brain Sciences, vol. 22, no. 4, 1999.
-  Nick Hawes, Aaron Sloman, Jeremy Wyatt, Michael Zillich, Henrik Jacobsson, Geert-Jan M. Kruijff, Michael Brenner, Gregor Berginc & Danijel Skocaj.
Towards an Integrated Robot with Multiple Cognitive Functions.
In AAI, pages 1548–1553. AAAI Press, 2007.



Bibliography II

-  P. Knoeferle & M.C. Crocker.
The coordinated interplay of scene, utterance, and world knowledge : evidence from eye tracking.
Cognitive Science, 2006.
-  Geert-Jan M. Kruijff.
What makes a cognitive system understand situated dialogue ?, March 2006.
invited talk at the Department of Linguistics, University of Texas, Austin.
-  Geert-Jan M. Kruijff.
How to make talking robots, August 2007.
IEEE RO-MAN 07 Tutorial, Jeju Island (Korea).



Bibliography III



George Lakoff.

Women, fire and dangerous things : what categories reveal about the mind.

University of Chicago Press, Chicago, 1987.



R. Lippmann.

Speech recognition by machines and humans.

Speech Communication, vol. 22, no. 1, pages 1–16, 1997.