Robust processing of spoken situated dialogue

Pierre Lison
Cognitive Systems @ Language Technology Lab
German Research Centre for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany.
[pierrel@coli.uni-sb.de]

18. Tagung der Computerlinguistik-Studierenden (TaCoS),
Universität Potsdam, Germany.
June 2008
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 title is: 

« *Robust processing of spoken situated dialogue* ».

Practically, I tried to improve the performance of a dialogue system incorporated in a robotic platform, by making it more **robust** to ill-formed utterances.

And this at various processing stages, from speech recognition up to semantic interpretation.
For this talk, I’ll focus on one particular technique I developed, which pertains to speech recognition.

The technique in question is a type of context-sensitive speech recognition which relies on the visual and discourse contexts to prime recognition.

The system is fully implemented as part of a cognitive architecture for mobile robots interacting with humans to perform a variety of service-oriented tasks.
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
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

3 Conclusions

4 Bibliography
Introduction

Outline

Outline of the talk

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

2 Approach
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

3 Conclusions

4 Bibliography

Pierre Lison
Robust processing of spoken situated dialogue
Outline of the talk

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

2 Approach
   • The issue
   • Proposed solution
   • Implementation
   • Evaluation

3 Conclusions

4 Bibliography
Outline of the talk

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

2. Approach
   - The issue
   - Proposed solution
   - Implementation
   - Evaluation

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:

- Robust processing of spoken situated dialogue
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.
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 [Hawes 07].
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.
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.
Schematic view of the CoSy architecture we are currently developing for spoken dialogue comprehension:

- **Speech recognition**
  - Automatic speech recognition with Nuance v8.5
  - Dynamic update of statistical language model

- **Recognition result**
  - Packed logical form

- **Incremental parsing**
  - Incremental chart parsing with COG grammar
  - Padding and pruning of logical forms

- **Dialogue interpretation**
  - Dialogue structure update (SDRIF)
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* and why it is so difficult to achieve in real-world environments.

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* and why it is so difficult to achieve in real-world environments.

- I’ll now present my own (modest) contribution to the current research in the field.
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* and why it is so difficult to achieve in real-world environments.

I’ll now present my own (modest) contribution to the current research in the field.
What have we seen so far?

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* and why it is so difficult to achieve in real-world environments.

I’ll now present my own (modest) contribution to the current research in the field.
The first step in comprehending spoken dialogue is *automatic speech recognition* \([\text{ASR}]\).

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 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.
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].
Practically, we use two main sources of information:

1. objects in the perceived \textit{visual scene};
2. linguistic expressions in the \textit{dialogue history}.

These objects are then ranked according to their \textit{saliency}, and integrated into a \textit{cross-modal salience model}.

This salience model is then applied to dynamically compute \textit{lexical activations}, which are incorporated into the language model of the speech recogniser.
Lexical activation

- A *lexical activation network* 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.

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.
Let’s imagine we are in the lab with the robot. There is a big red ball in front of him (= high saliency).

The red ball is perceived by the robot sensors (camera, laser scanner, etc.), and recognised as a “red ball”.

In the robot’s knowledge base, the “red ball” object is associated to words like “ball” like “round”, “pick up”, etc.

As a final step, we adapt the language model included in the speech recogniser to increase the probability of hearing these words.
Language Modeling

- The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg \max_W P(W|O, E)$$

$$= \arg \max_W P(O|W) \times \frac{P(W|E)}{\text{acoustic model} \times \text{salience-driven language model}}$$  (2)

- For a trigram language model, the probability of the word sequence $P(w_1^n|E)$ is:

$$P(w_1^n|E) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E)$$  (3)
The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg \max_W P(W|O, E)$$

$$= \arg \max_W P(O|W) \times \frac{P(W|E)}{P(O|W)}$$

(1)

For a trigram language model, the probability of the word sequence $P(w_1^n|E)$ is:

$$P(w_1^n|E) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E)$$

(2)
Language Modeling

The objective of the situated speech recognizer is to find the word sequence $W^*$ which has the highest probability given the observed speech signal $O$ and a set $E$ of salient objects:

$$W^* = \arg \max_W P(W|O, E)$$

$$= \arg \max_W P(O|W) \times \frac{P(W|E)}{P(O|W) \times P(W|E)}$$

For a trigram language model, the probability of the word sequence $P(w_1^n|E)$ is:

$$P(w_1^n|E) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}; E)$$
Our language model is class-based, so it can be further decomposed into word-class and class transition probabilities:

- Word-class probabilities: \( P(w_i|c_i) \)
- Class-transition probabilities: \( P(c_i|c_{i-1}, c_{i-2}) \).

The class transition probabilities reflect the language syntax - we assume they are independent of salient objects. The word-class probabilities, however, *do* depend on context.

The probability of encountering the word \( w_i \) in the sequence \( w_1^n \) using a class-based trigram model is therefore defined as:

\[
P(w_i|w_{i-1}w_{i-2}; E) = \underbrace{P(w_i|c_i; E)}_{\text{word-class prob.}} \times \underbrace{P(c_i|c_{i-1}, c_{i-2})}_{\text{class transition prob.}} \tag{4}
\]
We finally define the word-class probabilities $P(w_i|c_i; \mathbf{E})$:

$$P(w_i|c_i; \mathbf{E}) = \sum_{e_k \in \mathbf{E}} P(w_i|c_i; e_k) \times P(e_k)$$

with $e_k \in \mathbf{E}$ representing a specific salient object.

To compute $P(w_i|c_i; e_k)$, we use the **lexical activation network** specified for $e_k$.

To put it simply, we **increase** the probability of words with are activated by $e_k$ and **decrease** the probability of the others.

The probabilities are **dynamically updated** as the environment and the dialogue evolves and incorporated into the language model at **runtime**.
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.
<table>
<thead>
<tr>
<th>Word Error Rate [WER]</th>
<th>Classical LM</th>
<th>Salience-driven LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>vocab. size ≃ 200 words</td>
<td>25.04 % (NBest 3 : 20.72 %)</td>
<td>24.22 % (NBest 3 : 19.97 %)</td>
</tr>
<tr>
<td>vocab. size ≃ 400 words</td>
<td>26.68 % (NBest 3 : 21.98 %)</td>
<td>23.85 % (NBest 3 : 19.97 %)</td>
</tr>
<tr>
<td>vocab. size ≃ 600 words</td>
<td>28.61 % (NBest 3 : 24.59 %)</td>
<td>23.99 % (NBest 3 : 20.27 %)</td>
</tr>
</tbody>
</table>

**Tab.**: Comparative results of recognition performance
Conclusions

- In this talk, I explained how we could develop robots endowed with *communicative abilities*, i.e. 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 one particular technique devised to alleviate this problem, which uses the the *situated context* to improve the speech recognition performance.
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 one particular technique devised to alleviate this problem, which uses the situated context to improve the speech recognition performance.
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 one particular technique devised to alleviate this problem, which uses the the *situated context* to improve the speech recognition performance.
Thank you for your attention!!

⇒ Questions, comments?

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

Bibliography II

P. Knoeferle & M.C. Crocker.  
*The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking.*  

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.  
IEEE RO-MAN 07 Tutorial, Jeju Island (Korea).