Robust Processing of Spoken Situated Dialogue

Master thesis

Pierre Lison

Cognitive Systems @ Language Technology Lab

German Research Centre for Artificial Intelligence (DFKI GmbH)

[pierre.lison@dfki.de]

Supervisors: Dr. ir. Geert-Jan Kruijff
Prof. Dr. Hans Uszkoreit

Universität des Saarlandes, Germany.
November 2008.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

CoSy Project
"Cognitive systems for cognitive assistants"
EU FP6 Integrated project

Outline of the talk

1. Introduction
   - The issues
   - Our approach in brief
2. Background
   - Spoken dialogue
   - Software architecture for HRI
3. Approach
   - Step 1 : Situated speech recognition
   - Step 2 : Grammar relaxation
   - Step 3 : Discriminative parse selection
   - Experimental evaluation
4. Conclusions

Talking robots ?

- Our long-term aim :

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

- **Research goal** : building robots which are able to understand (and produce) *situated, spoken dialogue.*
- **Question** : How can we achieve that, given the current limitations of NLP technology ?

Talking robots ?

- Our long-term aim :

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

- **Research goal** : building robots which are able to understand (and produce) *situated, spoken dialogue*.
- **Question** : How can we achieve that, given the current limitations of NLP technology ?

Issues in spoken dialogue understanding

- Dialogue systems typically suffer from a lack of *robustness* and *adaptivity*.
- **Four issues** of particular importance :
  1. Difficulty of accommodating *spoken language phenomena* (disfluencies, fragments, etc.) in the dialogue system ;
  2. Pervasiveness of speech recognition errors ;
  3. Ambiguities arising at all processing levels ;
  4. Extra-grammaticality.

Issues in spoken dialogue understanding

- Dialogue systems typically suffer from a lack of *robustness* and *adaptivity*.

- **Four issues** of particular importance :

  1. Difficulty of accommodating *spoken language phenomena* (disfluencies, fragments, etc.) in the dialogue system ;
  2. Pervasiveness of speech recognition errors ;
  3. Ambiguities arising at all processing levels ;
  4. Extra-grammaticality.

Issues in spoken dialogue understanding

- Dialogue systems typically suffer from a lack of *robustness* and *adaptivity*.

- **Four issues** of particular importance :

  1. Difficulty of accommodating *spoken language phenomena* (disfluencies, fragments, etc.) in the dialogue system ;
  2. Pervasiveness of speech recognition errors ;
  3. Ambiguities arising at all processing levels ;
  4. Extra-grammaticality.

Issues in spoken dialogue understanding

- Dialogue systems typically suffer from a lack of *robustness* and *adaptivity*.
- **Four issues** of particular importance :
  1. Difficulty of accommodating *spoken language phenomena* (disfluencies, fragments, etc.) in the dialogue system ;
  2. Pervasiveness of speech recognition errors ;
  3. Ambiguities arising at all processing levels ;
  4. Extra-grammaticality.

Issues in spoken dialogue understanding

- Dialogue systems typically suffer from a lack of *robustness* and *adaptivity*.
- **Four issues** of particular importance :
  1. Difficulty of accommodating *spoken language phenomena* (disfluencies, fragments, etc.) in the dialogue system ;
  2. Pervasiveness of speech recognition errors ;
  3. Ambiguities arising at all processing levels ;
  4. Extra-grammaticality.

What the thesis is about

- We present an **integrated approach** for addressing these questions, in the context of domain-specific dialogues for human-robot interaction.

- The approach is *fully implemented*, and integrated in a cognitive architecture for autonomous robots.

- We performed an extensive *evaluation* of our approach.

  $\Rightarrow$ The empirical results we obtained demonstrate **very significant improvements** both in *robustness* and in *accuracy* compared to the baseline.

What the thesis is about

- We present an **integrated approach** for addressing these questions, in the context of domain-specific dialogues for human-robot interaction.
- The approach is *fully implemented*, and integrated in a cognitive architecture for autonomous robots.
- We performed an extensive *evaluation* of our approach.

  ⇒ The empirical results we obtained demonstrate **very significant improvements** both in *robustness* and in *accuracy* compared to the baseline.

What the thesis is about

- We present an **integrated approach** for addressing these questions, in the context of domain-specific dialogues for human-robot interaction.
- The approach is *fully implemented*, and integrated in a cognitive architecture for autonomous robots.
- We performed an extensive *evaluation* of our approach.

  ⇒ The empirical results we obtained demonstrate **very significant improvements** both in *robustness* and in *accuracy* compared to the baseline.

The strategy in three steps

1. Improve the performance of speech recognition by exploiting **contextual knowledge** about the *environment* and the *dialogue state*.

2. Allow for a **controlled relaxation** of the grammatical constraints to account for *spoken dialogue phenomena* and *speech recognition errors*.

3. Finally, apply a **discriminative model** on the resulting set of interpretations, in order to select the most likely one given the context.

The strategy in three steps

1. Improve the performance of speech recognition by exploiting **contextual knowledge** about the *environment* and the *dialogue state*.

2. Allow for a **controlled relaxation** of the grammatical constraints to account for *spoken dialogue phenomena* and *speech recognition errors*.

3. Finally, apply a **discriminative model** on the resulting set of interpretations, in order to select the most likely one given the context.

The strategy in three steps

1. Improve the performance of speech recognition by exploiting **contextual knowledge** about the *environment* and the *dialogue state*.

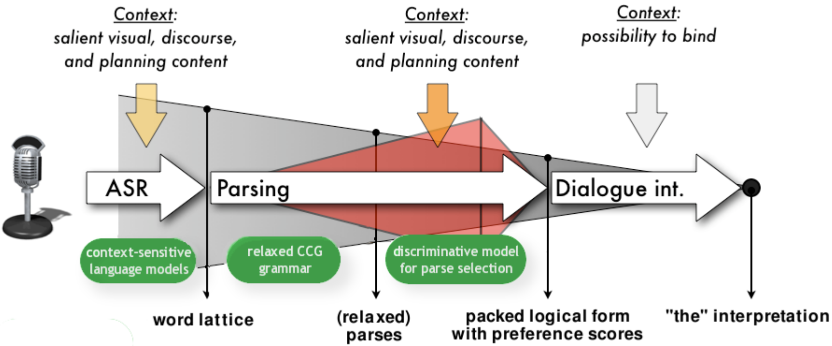
2. Allow for a **controlled relaxation** of the grammatical constraints to account for *spoken dialogue phenomena* and *speech recognition error*s.

3. Finally, apply a **discriminative model** on the resulting set of interpretations, in order to select the most likely one given the context.

The strategy in three steps

1. Improve the performance of speech recognition by exploiting **contextual knowledge** about the *environment* and the *dialogue state*.
2. Allow for a **controlled relaxation** of the grammatical constraints to account for *spoken dialogue phenomena* and *speech recognition errors*.
3. Finally, apply a **discriminative model** on the resulting set of interpretations, in order to select the most likely one given the context.

The strategy, graphically



Spoken dialogue

Different levels of processing :

- **Auditory** : speech recognition
- **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, etc.
  "Discourse interprets utterance meaning relative to the context, establishing how it contributes to furthering the discourse"

Spoken dialogue

Different levels of processing :

- **Auditory** : speech recognition
- **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, etc.

  "Discourse interprets utterance meaning relative to the 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
- **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

- Extract from a corpus of task-oriented spoken dialogue : *The Apollo Lunar Surface Journal*. [Audio file]

Example

Parker : That's all we need. Go ahead and park on your 045 <**okay**>. We'll give you an update when you're done.
Cernan : Jack is **[it]** worth coming right there ?
Schmitt : **err** looks like a pretty **go/** good location.
Cernan : okay.
Schmitt : We can sample the rim materials of this crater. **(Pause)** Bob, I'm at the **uh** south **uh** let's say east-southeast rim of a, **oh**, 30-meter crater - **err** in the light mantle, of course - up on the **uh** Scarp and maybe 300...**(correcting himself)** **err** 200 meters from the **uh** rim of Lara in **(inaudible)** northeast direction.

Software architecture

- Software architectures for "intelligent" robots are typically composed of several *distributed* and *cooperating* subsystems, such as :
  - communication ;
  - vision, perception ;
  - navigation and manipulation skills ;
  - deliberative processes (for planning, learning, reasoning).
- Our approach has been implemented as part of a *distributed cognitive architecture* for autonomous robots.
- In this presentation we focus only on the **communication** subarchitecture.

Software architecture

- Software architectures for “intelligent” robots are typically composed of several *distributed* and *cooperating* subsystems, such as :
  - communication ;
  - vision, perception ;
  - navigation and manipulation skills ;
  - deliberative processes (for planning, learning, reasoning).
- Our approach has been implemented as part of a *distributed cognitive architecture* for autonomous robots.
- In this presentation we focus only on the **communication** subarchitecture.

Communication subarchitecture

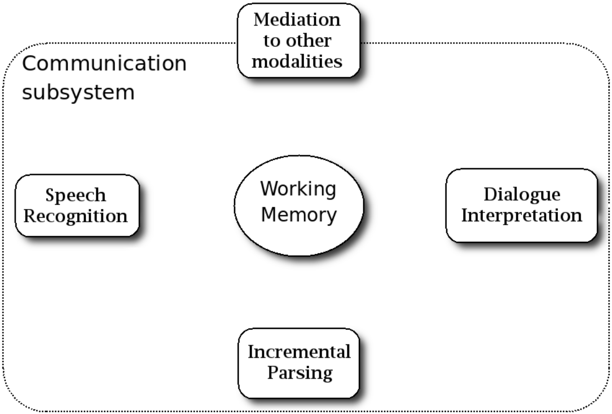


FIG.: Spoken dialogue comprehension

Communication subarchitecture

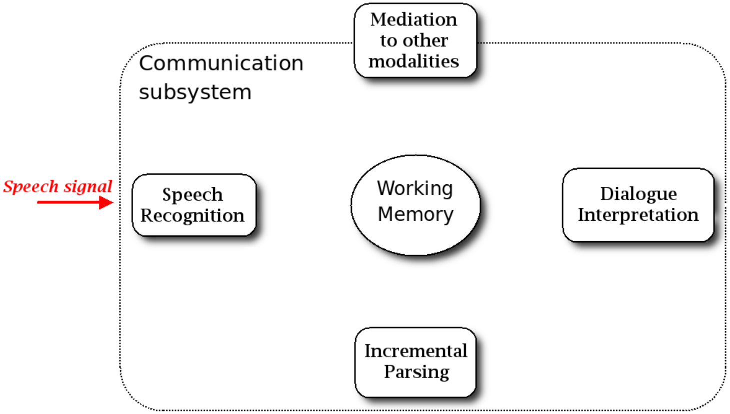


FIG.: Spoken dialogue comprehension : step 1

Communication subarchitecture

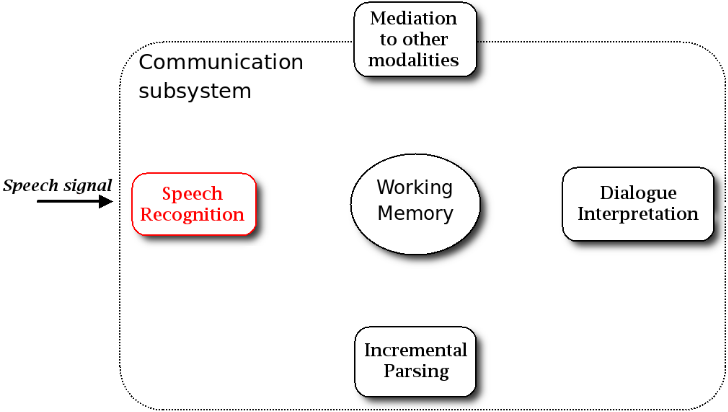


FIG.: Spoken dialogue comprehension : step 1

Communication subarchitecture

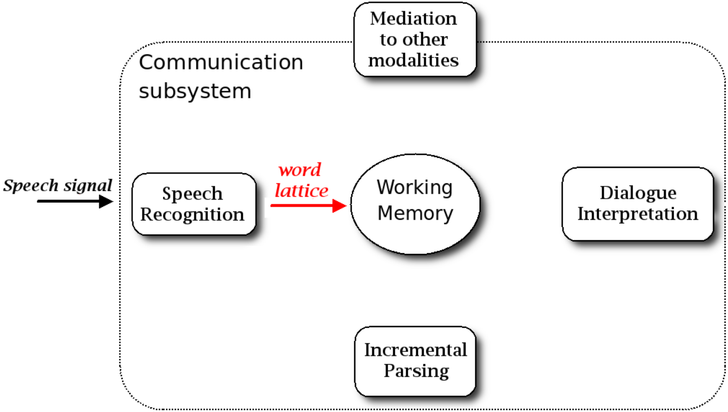


FIG.: Spoken dialogue comprehension : step 1

Communication subarchitecture

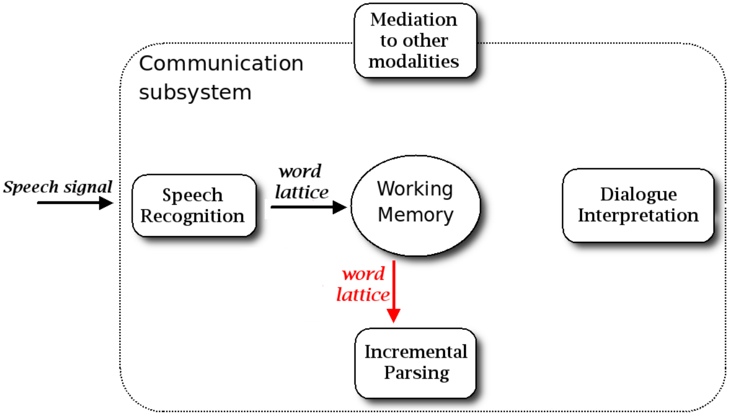


FIG.: Spoken dialogue comprehension : step 2

Communication subarchitecture

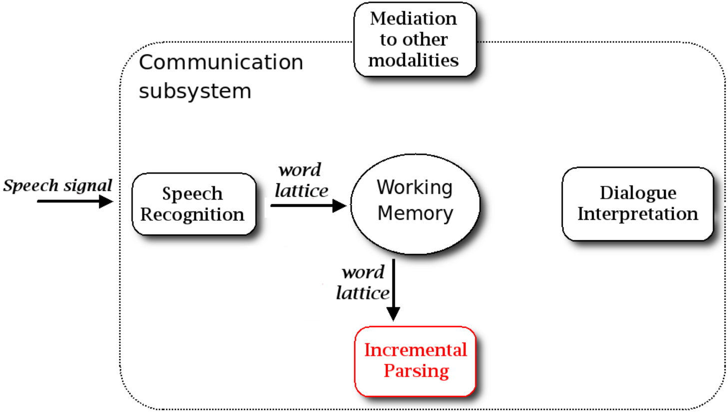


FIG.: Spoken dialogue comprehension : step 2

Communication subarchitecture

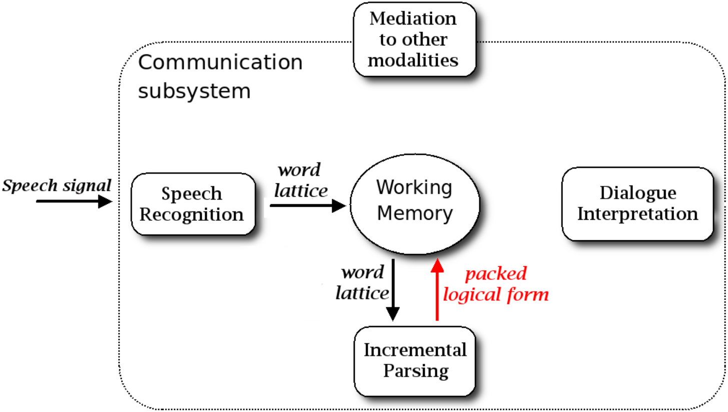


FIG.: Spoken dialogue comprehension : step 2

Communication subarchitecture

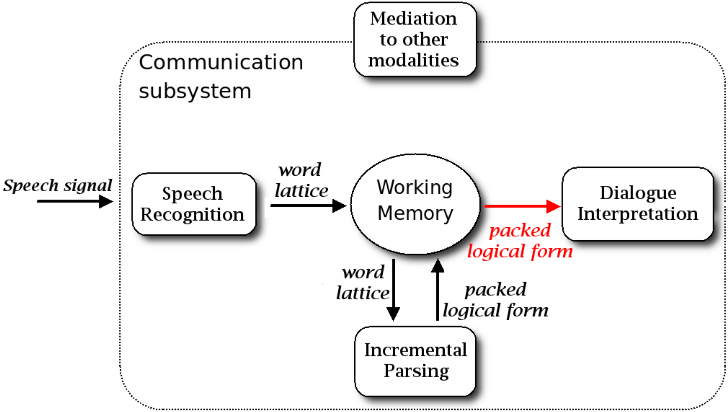


FIG.: Spoken dialogue comprehension : step 3

Communication subarchitecture

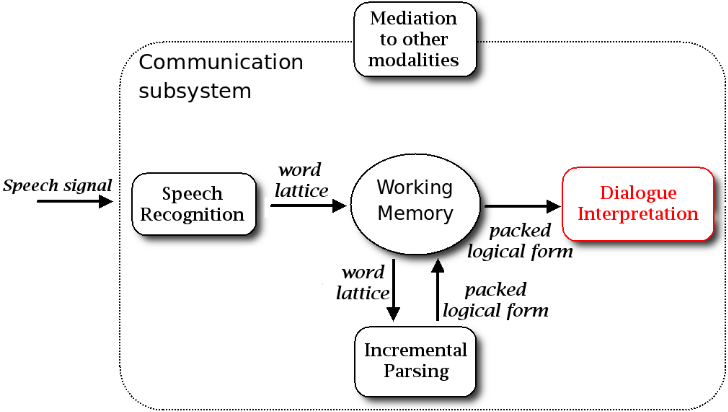


FIG.: Spoken dialogue comprehension : step 3

Communication subarchitecture

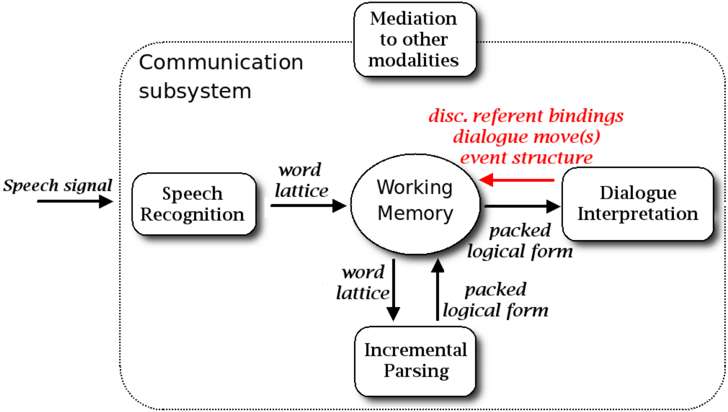


FIG.: Spoken dialogue comprehension : step 3

Communication subarchitecture

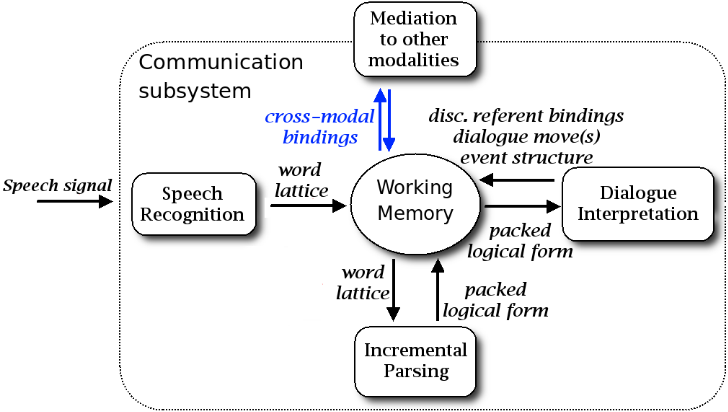


FIG.: Spoken dialogue comprehension : cross-modality

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*.

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*.

Proposed solution

- The **intuition** underlying our approach : *use context !*

- More precisely, we *prime* the utterance recognition by exploiting information about
  1. The salient entities in the situated visual environment ;
  2. The dialogue state.

- Our **claim** : for HRI, the speech recognition performance can be *significantly enhanced* by using contextual knowledge.

Proposed solution

- The **intuition** underlying our approach : *use context !*
- More precisely, we *prime* the utterance recognition by exploiting information about
  1. The salient entities in the situated visual environment ;
  2. The dialogue state.
- Our **claim** : for HRI, the speech recognition performance can be *significantly enhanced* by using contextual knowledge.

Implementation

- 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 **saliency**, 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.

Implementation

- 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 **saliency**, 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

- A **lexical activation network** lists, for each possible salient entity, the set of words activated by it.
- In other words, it specifies *the words which are likely to be heard* when the given entity is present in the environment.
- 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.
- $\Rightarrow$ The evaluation results showed a significant reduction of the word error rate compared to the baseline (-16.1% compared to the baseline, p-value is $1.9 \times 10 - 3$).

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Robust parsing of spoken inputs

- Parsing spoken inputs is a difficult task
- The parser must be made robust to *ill-formed* and *misrecognised* inputs
- Three broad families of techniques can be used :
  - Shallow or partial parsing (concept spotting) ;
  - (pure) statistical approaches (HMMs, stochastic parsers) ;
  - Controlled relaxation of grammar rules.
- Our approach is based on **grammar relaxation**.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar
- The new rules are :
  - **New type-shifting rules**
    ⇒ to account for missing words ;
  - **“Paradigmatic heap” rules**
    ⇒ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    ⇒ to be able to combine discourse units ;
  - And **ASR correction rules**
    ⇒ to correct misrecognised words.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar
- The new rules are :
  - **New type-shifting rules**
    ⇒ to account for missing words ;
  - **"Paradigmatic heap" rules**
    ⇒ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    ⇒ to be able to combine discourse units ;
  - And **ASR correction rules**
    ⇒ to correct misrecognised words.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar
- The new rules are :
  - **New type-shifting rules**
    $\Rightarrow$ to account for missing words ;
  - **"Paradigmatic heap" rules**
    $\Rightarrow$ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    $\Rightarrow$ to be able to combine discourse units ;
  - And **ASR correction rules**
    $\Rightarrow$ to correct misrecognised words.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar
- The new rules are :
  - **New type-shifting rules**
    $\Rightarrow$ to account for missing words ;
  - **"Paradigmatic heap" rules**
    $\Rightarrow$ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    $\Rightarrow$ to be able to combine discourse units ;
  - And **ASR correction rules**
    $\Rightarrow$ to correct misrecognised words.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar

- The new rules are :
  - **New type-shifting rules**
    ⇒ to account for missing words ;
  - **“Paradigmatic heap” rules**
    ⇒ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    ⇒ to be able to combine discourse units ;
  - And **ASR correction rules**
    ⇒ to correct misrecognised words.

Implementation

- Practically, the relaxation is realised by introducing **non-standard CCG combinators** into the grammar
- The new rules are :
  - **New type-shifting rules**
    $\Rightarrow$ to account for missing words ;
  - **“Paradigmatic heap” rules**
    $\Rightarrow$ to account for syntactic disfluencies ;
  - **Discourse-level composition rules**
    $\Rightarrow$ to be able to combine discourse units ;
  - And **ASR correction rules**
    $\Rightarrow$ to correct misrecognised words.

A simple example

$$\dfrac{\dfrac{\dfrac{\text{pick}}{\text{s/particle/np}} \quad \dfrac{\dfrac{\text{cup}}{\text{up}}\ corr}{\text{particle}}}{\text{s/np}}> \quad \dfrac{\dfrac{\text{the}}{\text{np/n}} \quad \dfrac{\text{ball}}{\text{n}}}{\text{np}}>}{\text{s}}>$$

Fig.: CCG derivation of "pick cup the ball".

Discriminative models

- The set of interpretations resulting from the parsing operation can be quite large.

- Why ?

  1. Multiple recognition hypotheses from the ASR ;
  2. Controlled relaxation of grammatical constraints ;
  3. And finally, language is inherently ambiguous, and spoken dialogue is no exception !

- We need a mechanism which filters out unlikely interpretations and only keeps the good one(s).
  $\Rightarrow$ Integration of a **discriminative model** for parse selection

Discriminative models

- The set of interpretations resulting from the parsing operation can be quite large.

- Why ?

  1. Multiple recognition hypotheses from the ASR ;
  2. Controlled relaxation of grammatical constraints ;
  3. And finally, language is inherently ambiguous, and spoken dialogue is no exception !

- We need a mechanism which filters out unlikely interpretations and only keeps the good one(s).
  $\Rightarrow$ Integration of a **discriminative model** for parse selection

Discriminative models

- The set of interpretations resulting from the parsing operation can be quite large.

- Why ?

  1. Multiple recognition hypotheses from the ASR ;
  2. Controlled relaxation of grammatical constraints ;
  3. And finally, language is inherently ambiguous, and spoken dialogue is no exception !

- We need a mechanism which filters out unlikely interpretations and only keeps the good one(s).
  $\Rightarrow$ Integration of a **discriminative model** for parse selection

Discriminative models

- The set of interpretations resulting from the parsing operation can be quite large.
- Why ?
  1. Multiple recognition hypotheses from the ASR ;
  2. Controlled relaxation of grammatical constraints ;
  3. And finally, language is inherently ambiguous, and spoken dialogue is no exception !
- We need a mechanism which filters out unlikely interpretations and only keeps the good one(s).
  $\Rightarrow$ Integration of a **discriminative model** for parse selection

Parse selection

- The task is defined as a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ where the domain $\mathcal{X}$ is the set of possible inputs (in our case, $\mathcal{X}$ is the set of possible *word lattices*), and $\mathcal{Y}$ the set of parses.

- The function F, mapping a word lattice to its most likely parse, is then defined as :

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \ \mathbf{w}^T \cdot \mathbf{f}(x, y) \tag{1}$$

where $\mathbf{w}^T \cdot \mathbf{f}(x, y)$ is the inner product $\sum_{s=1}^{d} w_s \ f_s(x, y)$, and can be seen as a measure of the "quality" of the parse.

Parse selection (cont'd)

- We assume :
  1. A function $\mathbf{GEN}(x)$ which enumerates all possible parses for an input x. In our case, this function represents the parses of x which are admissible according to the CCG grammar.
  2. A *d*-dimensional feature vector $\mathbf{f}(x, y) \in \Re^d$, representing specific features of the pair (x, y). It can include various acoustic, syntactic, semantic or contextual features which can be relevant in discriminating the parses
  3. A parameter vector $\mathbf{w} \in \Re^d$.

- Given the parameters $\mathbf{w}$, the optimal parse of a given utterance x is determined by enumerating all parses generated by the grammar, extracting their features, computing the inner product $\mathbf{w}^T \cdot \mathbf{f}(x, y)$, and selecting the highest-scoring parse.

Discriminative models : learning

- How do we learn the parameters $\mathbf{w}$?

- We use a well-known algorithm from machine learning : a **perceptron**.

- The perceptron algorithm has proven to be very efficient and accurate for the task of parse selection

- **Problem** :we don't have any annotated corpora for our domain at our disposal
  ⇒ **Solution** : automatic *generation* of training examples from a small domain-specific grammar.

Discriminative models : learning

- How do we learn the parameters $\mathbf{w}$?
- We use a well-known algorithm from machine learning : a **perceptron**.
- The perceptron algorithm has proven to be very efficient and accurate for the task of parse selection
- **Problem** :we don't have any annotated corpora for our domain at our disposal
  $\Rightarrow$ **Solution** : automatic *generation* of training examples from a small domain-specific grammar.

Discriminative models : learning

- How do we learn the parameters $\mathbf{w}$?
- We use a well-known algorithm from machine learning : a **perceptron**.
- The perceptron algorithm has proven to be very efficient and accurate for the task of parse selection
- **Problem** :we don't have any annotated corpora for our domain at our disposal

  $\Rightarrow$ **Solution** : automatic *generation* of training examples from a small domain-specific grammar.

Discriminative models : learning

- How do we learn the parameters $\mathbf{w}$?
- We use a well-known algorithm from machine learning : a **perceptron**.
- The perceptron algorithm has proven to be very efficient and accurate for the task of parse selection
- **Problem** :we don't have any annotated corpora for our domain at our disposal
  $\Rightarrow$ **Solution** : automatic *generation* of training examples from a small domain-specific grammar.

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Discriminative models : features

- The accuracy of our discriminative model crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model
- That is, features which help *discriminating* the parses.
- They must also be relatively cheap to compute.
- In our model, the features are of four types :
  1. **semantic features** (substructures of the logical form) ;
  2. **syntactic features** (derivational history of the parse) ;
  3. **contextual features** (situated and dialogue contexts) ;
  4. and finally **ASR features** (scores from speech recognition).

Experimental setup

- The test suite is composed of 195 individual utterances collected during **Wizard-of-Oz experiments**.
- These experiments consist of a set of situated human-robot interactions relative to a shared visual scene.
- They were free both in form and content – they could include questions, assertions, commands, answers or clarifications.
- Interaction scenario : **Playmate** (object manipulation and visual learning with a robotic arm)

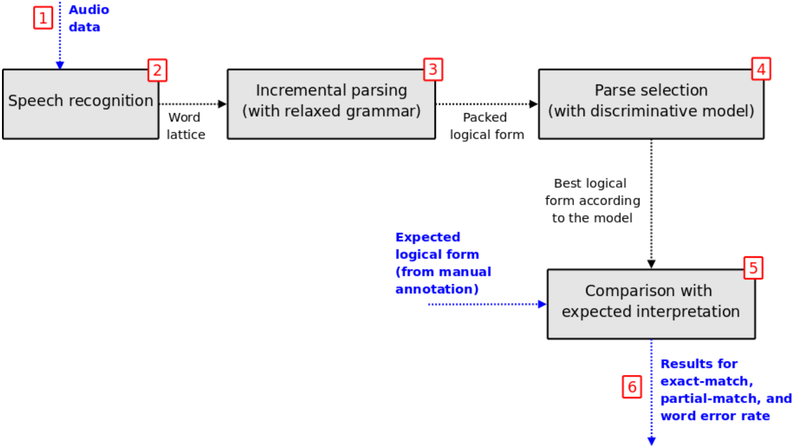
Experimental setup

- The test suite is composed of 195 individual utterances collected during **Wizard-of-Oz experiments**.
- These experiments consist of a set of situated human-robot interactions relative to a shared visual scene.
- They were free both in form and content – they could include questions, assertions, commands, answers or clarifications.
- Interaction scenario : **Playmate** (object manipulation and visual learning with a robotic arm)

Experimental setup (cont'd)

- Here is an example of interaction in the context of a playmate scenario : [Video file]
- The audio data resulting from these experiments was then manually *segmented*, *transcribed*, and *associated with a semantic annotation*.
- Our results are compared to a *baseline*.
- The **baseline** for our experiment is our dialogue comprehension system, but *without* grammar relaxation and discriminative models.

Experimental setup (2)



Evaluation results

	Precision	**Recall**	**$\mathbf{F_1}$-measure**
Baseline	40.9	45.2	43.0
Our approach	55.6	84.0	**66.9**

TAB.: Exact-match accuracy results (NBest 5 with all feats. activated)

	Precision	**Recall**	**$\mathbf{F_1}$-measure**
Baseline	86.2	56.2	68.0
Our approach	87.6	86.0	**86.8**

TAB.: Partial-match accuracy results (NBest 5 with all feats. activated)

- \+ significant decrease of the word error rate, going from **20.5** % for the baseline to **15.7** % with our approach. (p-value with t-tests is 0.036).

Evaluation results

	Precision	**Recall**	**$\mathbf{F_1}$-measure**
Baseline	40.9	45.2	43.0
Our approach	55.6	84.0	**66.9**

TAB.: Exact-match accuracy results (NBest 5 with all feats. activated)

	Precision	**Recall**	**$\mathbf{F_1}$-measure**
Baseline	86.2	56.2	68.0
Our approach	87.6	86.0	**86.8**

TAB.: Partial-match accuracy results (NBest 5 with all feats. activated)

- + significant decrease of the word error rate, going from **20.5** % for the baseline to **15.7** % with our approach. (p-value with t-tests is 0.036).

Evaluation results

	Precision	**Recall**	**F_1-measure**
Baseline	40.9	45.2	43.0
Our approach	55.6	84.0	**66.9**

TAB.: Exact-match accuracy results (NBest 5 with all feats. activated)

	Precision	**Recall**	**F_1-measure**
Baseline	86.2	56.2	68.0
Our approach	87.6	86.0	**86.8**

TAB.: Partial-match accuracy results (NBest 5 with all feats. activated)

- + significant decrease of the word error rate, going from **20.5** % for the baseline to **15.7** % with our approach. (p-value with t-tests is 0.036).

Our approach in brief

1. It is a **hybrid symbolic/statistical** approach
   - Combination of fined-grained linguistic resources with statistical models
   - Able to deliver both *deep* and *robust* dialogue processing
2. It is an **integrated** approach
   - Goes all the way from the speech signal up to the semantic & pragmatic interpretation
   - Interactions between various processing components
3. It is a **context-sensitive** approach
   - *Context* is used at every processing step to guide the processing
   - Both an *anticipation* tool and a *discrimination* tool

Our approach in brief

1. It is a **hybrid symbolic/statistical** approach
    - Combination of fined-grained linguistic resources with statistical models
    - Able to deliver both *deep* and *robust* dialogue processing

2. It is an **integrated** approach
    - Goes all the way from the speech signal up to the semantic & pragmatic interpretation
    - Interactions between various processing components

3. It is a **context-sensitive** approach
    - *Context* is used at every processing step to guide the processing
    - Both an *anticipation* tool and a *discrimination* tool

Our approach in brief

1. It is a **hybrid symbolic/statistical** approach
    - Combination of fined-grained linguistic resources with statistical models
    - Able to deliver both *deep* and *robust* dialogue processing
2. It is an **integrated** approach
    - Goes all the way from the speech signal up to the semantic & pragmatic interpretation
    - Interactions between various processing components
3. It is a **context-sensitive** approach
    - *Context* is used at every processing step to guide the processing
    - Both an *anticipation* tool and a *discrimination* tool

Our approach in brief

1. It is a **hybrid symbolic/statistical** approach
   - Combination of fined-grained linguistic resources with statistical models
   - Able to deliver both *deep* and *robust* dialogue processing
2. It is an **integrated** approach
   - Goes all the way from the speech signal up to the semantic & pragmatic interpretation
   - Interactions between various processing components
3. It is a **context-sensitive** approach
   - *Context* is used at every processing step to guide the processing
   - Both an *anticipation* tool and a *discrimination* tool

By way of conclusion

- Let's briefly recapitulate the main contributions of the thesis :

  1. A new model for **context-sensitive speech recognition**, which relies on the situated and dialogue context to dynamically adapt the ASR language model to the environment ;
  2. A new model for **robust parsing of spoken inputs**, based on a relaxed CCG grammar coupled with a discriminative model exploring a wide range of linguistic and contextual features.
  3. A fully working **implementation** for these two models, integrated into a cognitive architecture for autonomous robots. The implementation comes along with a complete set of training data and testing data.

By way of conclusion

- Let's briefly recapitulate the main contributions of the thesis :

  1. A new model for **context-sensitive speech recognition**, which relies on the situated and dialogue context to dynamically adapt the ASR language model to the environment ;
  2. A new model for **robust parsing of spoken inputs**, based on a relaxed CCG grammar coupled with a discriminative model exploring a wide range of linguistic and contextual features.
  3. A fully working **implementation** for these two models, integrated into a cognitive architecture for autonomous robots. The implementation comes along with a complete set of training data and testing data.

By way of conclusion

- Let's briefly recapitulate the main contributions of the thesis :

 1. A new model for **context-sensitive speech recognition**, which relies on the situated and dialogue context to dynamically adapt the ASR language model to the environment ;
 2. A new model for **robust parsing of spoken inputs**, based on a relaxed CCG grammar coupled with a discriminative model exploring a wide range of linguistic and contextual features.
 3. A fully working **implementation** for these two models, integrated into a cognitive architecture for autonomous robots. The implementation comes along with a complete set of training data and testing data.

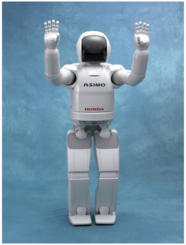
By way of conclusion

- Let's briefly recapitulate the main contributions of the thesis :

1. A new model for **context-sensitive speech recognition**, which relies on the situated and dialogue context to dynamically adapt the ASR language model to the environment ;
2. A new model for **robust parsing of spoken inputs**, based on a relaxed CCG grammar coupled with a discriminative model exploring a wide range of linguistic and contextual features.
3. A fully working **implementation** for these two models, integrated into a cognitive architecture for autonomous robots. The implementation comes along with a complete set of training data and testing data.

The end

Thank you for your attention ! !



⇒ **Questions, comments ?**

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