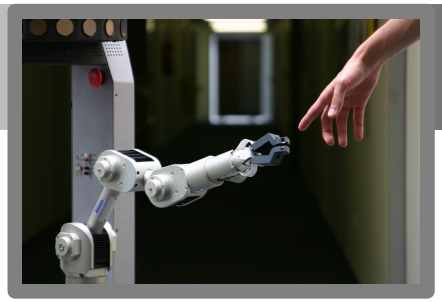
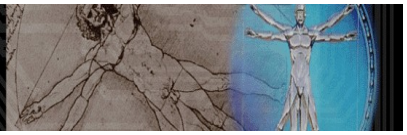


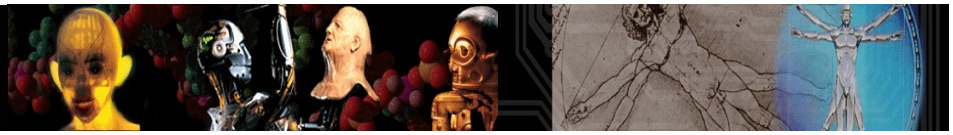


Incremental Processing of Spoken Dialogue For Human-Robot Interaction

Geert-Jan M. Kruijff
Pierre Lison

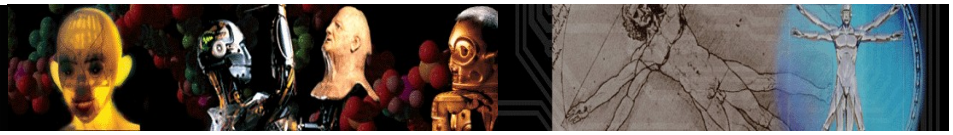
Language Technology Lab
DFKI GmbH, Saarbrücken
http://talkingrobots.dfki.de

Outline of the talk



- What is human-robot interaction?
- Overview of the approach
- Implementation
  - Situated speech recognition
  - Robust parsing of spoken dialogue
  - Improving efficiency with chart pruning
- Conclusion

Outline of the talk

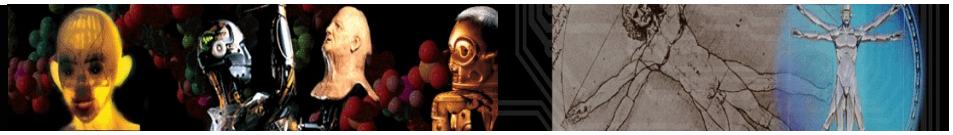


- **What is human-robot interaction?**
- Overview of the approach
- Implementation
  - Situated speech recognition
  - Robust parsing of spoken dialogue
  - Improving efficiency with chart pruning
- Conclusion

What is human-robot interaction?

- Communication in all its aspects
  - Verbal- and non-verbal behaviors,
  - including gesture, posture, affective display, ...
  - at various interaction ranges (proximal, distant),
  - with reference to varying spatio-temporal contexts
- HRI in this talk
  - Focus on spoken dialogue, proximal interaction, varying spatial contexts

Example scenario



- “Home tour„ scenario in an office environment
- Human-robot interaction with spoken dialogue
- Illustrates potential (miss-)communication problems
- The robot here remotely controlled in a Wizard-of-Oz fashion

Dialogue in HRI is (mostly) situated

Playing games on a table top...

Showing the robot around the “house“

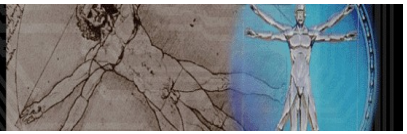
Teaching the robot about new objects

Describing what kind of object it should be looking for, in some other location,

And trying to ask someone how to get to that location.

- Situatedness of spoken dialogue in HRI
  - Spoken dialogue in our case is often referential to aspects of the environment
  - “The environment” may refer to *small-scale space*, e.g. a table top, an area we are in,
  - But may also concern *large-scale space*, going beyond what is currently visible.
- Using situatedness in processing dialogue
  - Complex problem of reference resolution
  - Priming comprehension on the basis of situated context

Some problems to tackle



- The “usual” for spoken dialogue in HRI
  - Just like human spoken dialogue, dialogue in HRI is rife with incomplete or incorrect utterances, self-corrections, etc.
  - Pervasiveness of speech recognition errors
  - Ambiguities can arise at all processing levels
  - Extra-grammaticality (“out-of-coverage”) in relatively free dialogue
- Draw inspiration from how humans process dialogue
  - In visually situated dialogue, there is a close coupling between how humans understand what they see, and what they hear
  - Priming effects: situated context (and temporal information, world knowledge) primes dialogue comprehension, incrementally

→ Incrementally use the situated context to select, refine, extend, complement the interpretations, and increase robustness

Disfluencies in spoken dialogue

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

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 ?

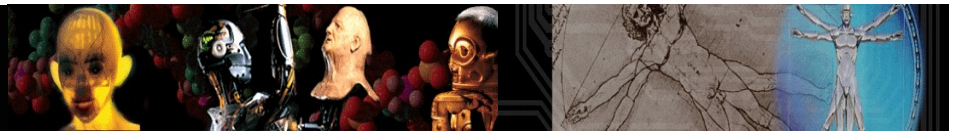
Schmitt : **err** looks like a pretty **go*l*** good location.

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.

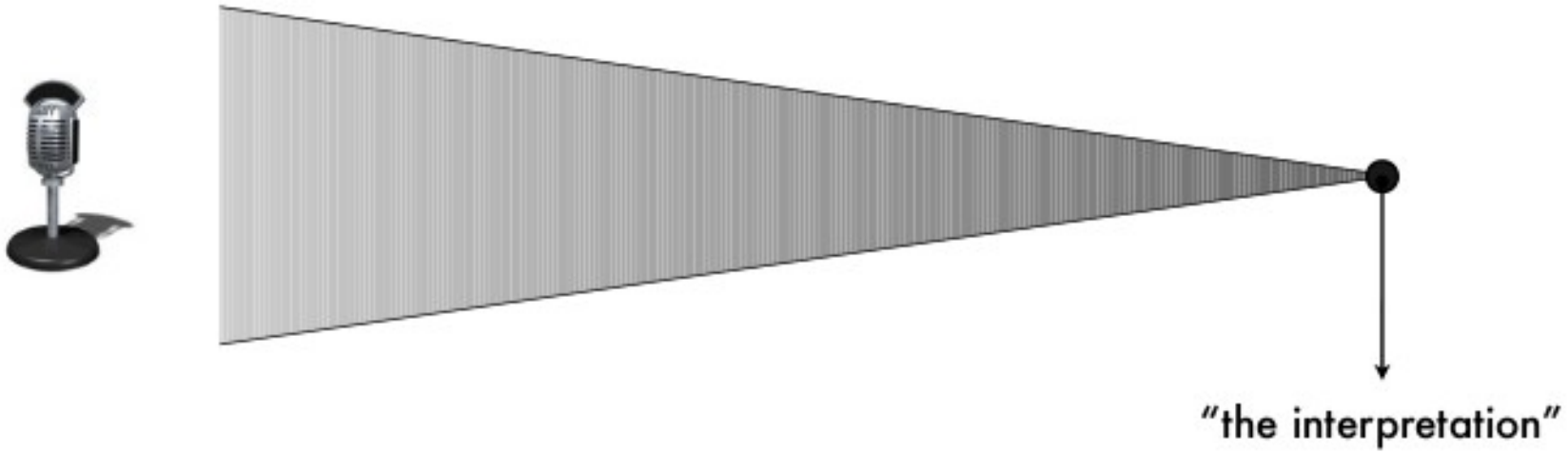
[Play sound file]

Outline of the talk

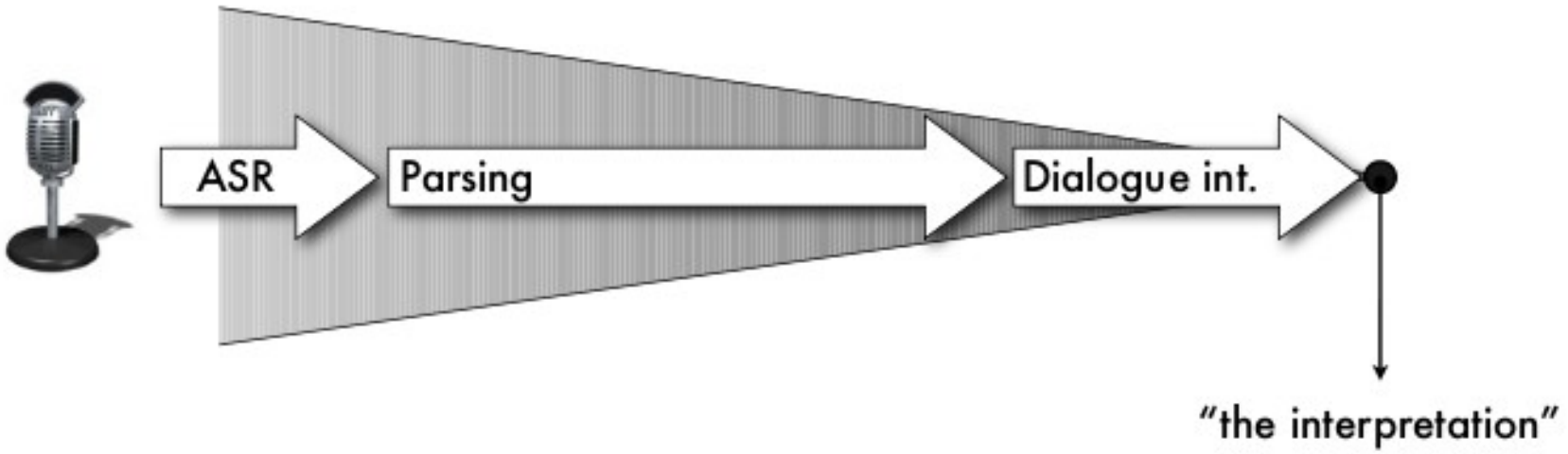


- What is human-robot interaction?
- **Overview of the approach**
- Implementation
  - Situated speech recognition
  - Robust parsing of spoken dialogue
  - Improving efficiency with chart pruning
- Conclusion

Overview of the approach

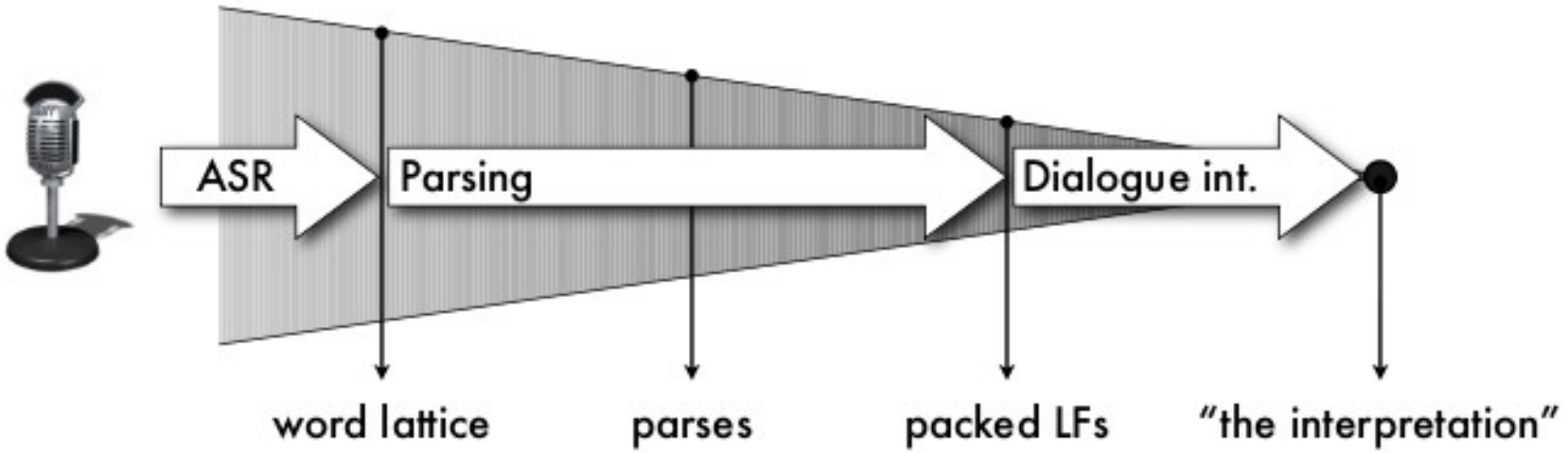


Overview of the approach



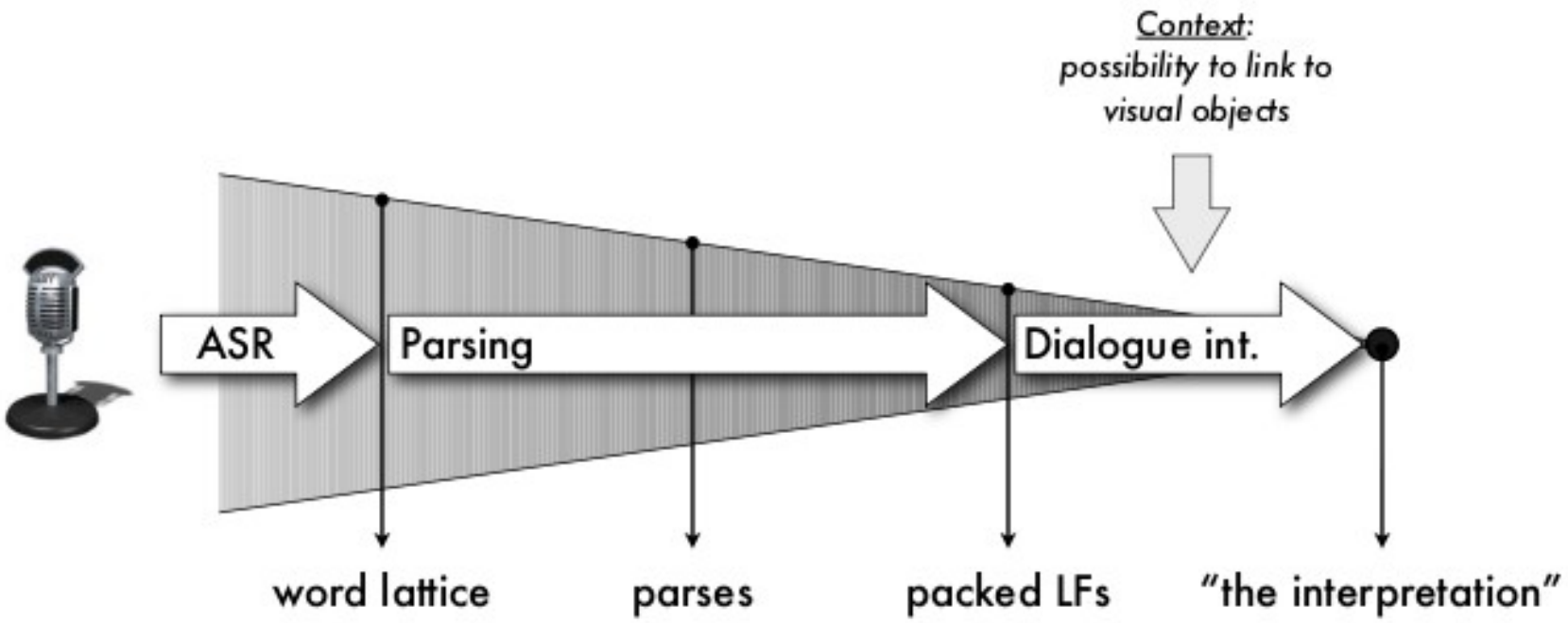
- Speech recognition with off-the-shelf ASR system
  - Language model is a class-based trigram statistical model
- Incremental parsing with Combinatory Categorial Grammar
- Dialogue interpretation tasks: reference resolution, dialogue move recognition, event structure interpretation

Overview of the approach



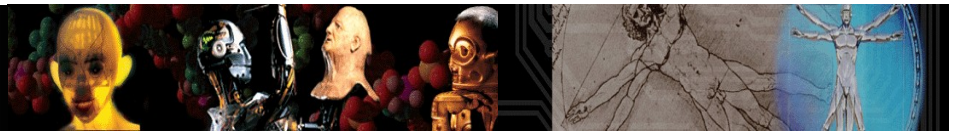
- Speech recognition outputs a *word lattice*
  - Word lattice = set of alternative recognition hypotheses compacted in a directed graph
- The CCG parser takes a word lattice as input and outputs packed logical forms, expressed in a hybrid logic formalism (HLDS)
  - Logical forms are ontologically rich, relational structures
- Dialogue interpretation based on a SDRT-like dialogue structure

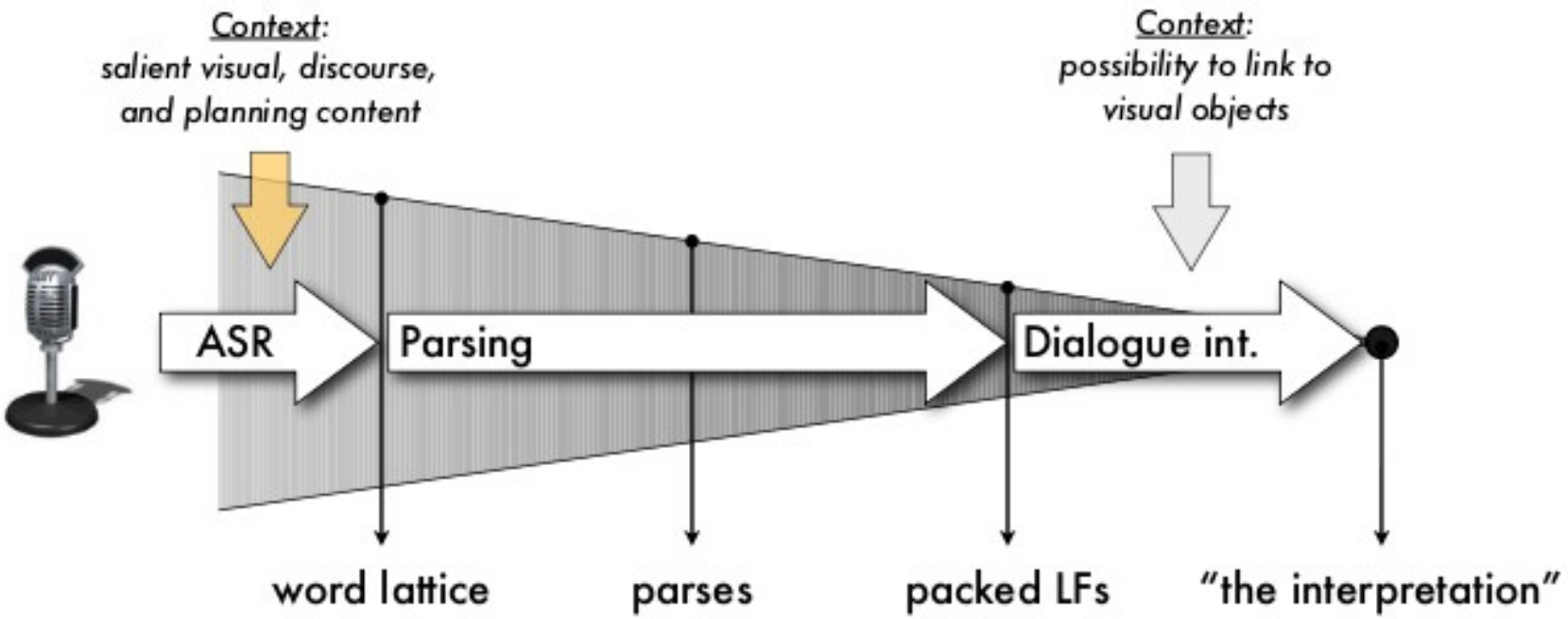
Overview of the approach



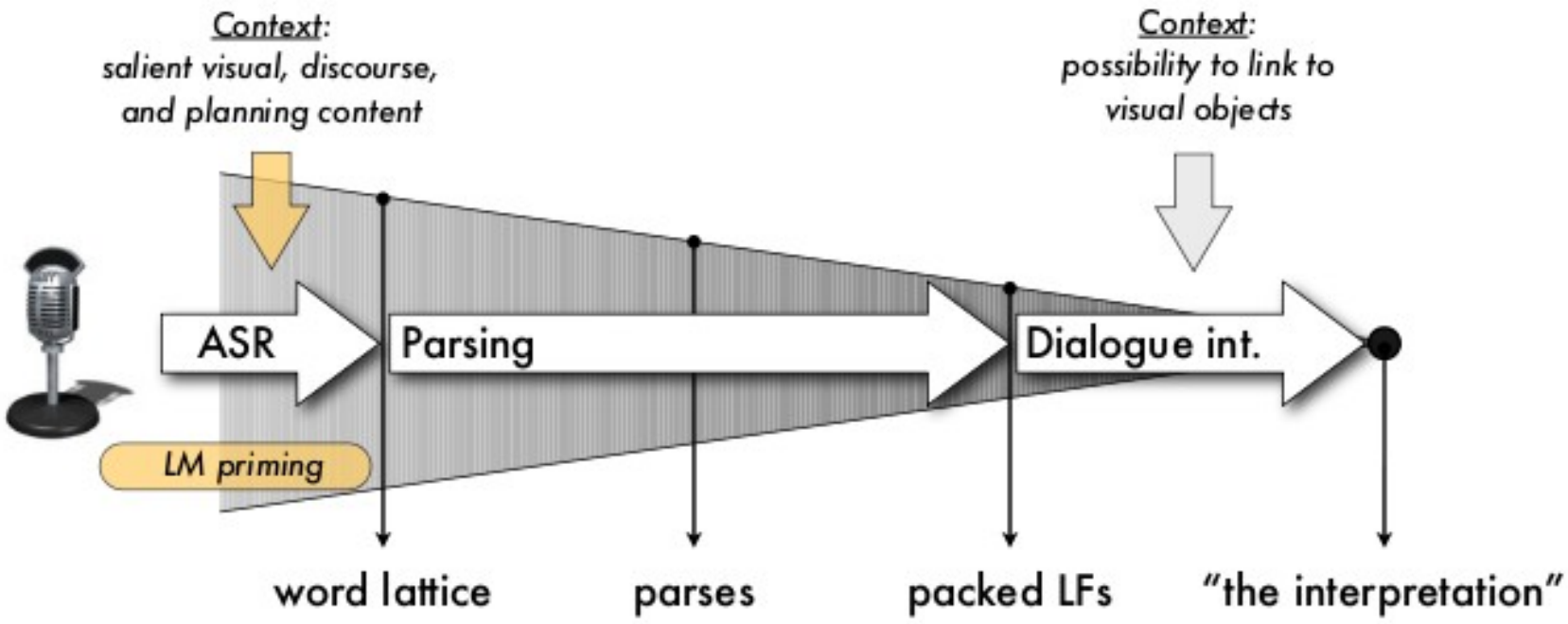
- Linguistic interpretations must be associated with extra-linguistic knowledge about the environment
  - Dialogue needs to connect with other modalities like vision, spatial reasoning or planning.
- A specific module, called the "binder", is responsible for this cross-modal information binding
  - Ontology-based *mediation* accross modalities

Overview of the approach



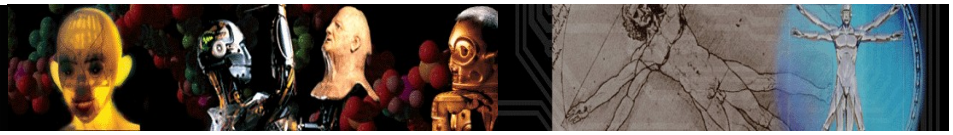


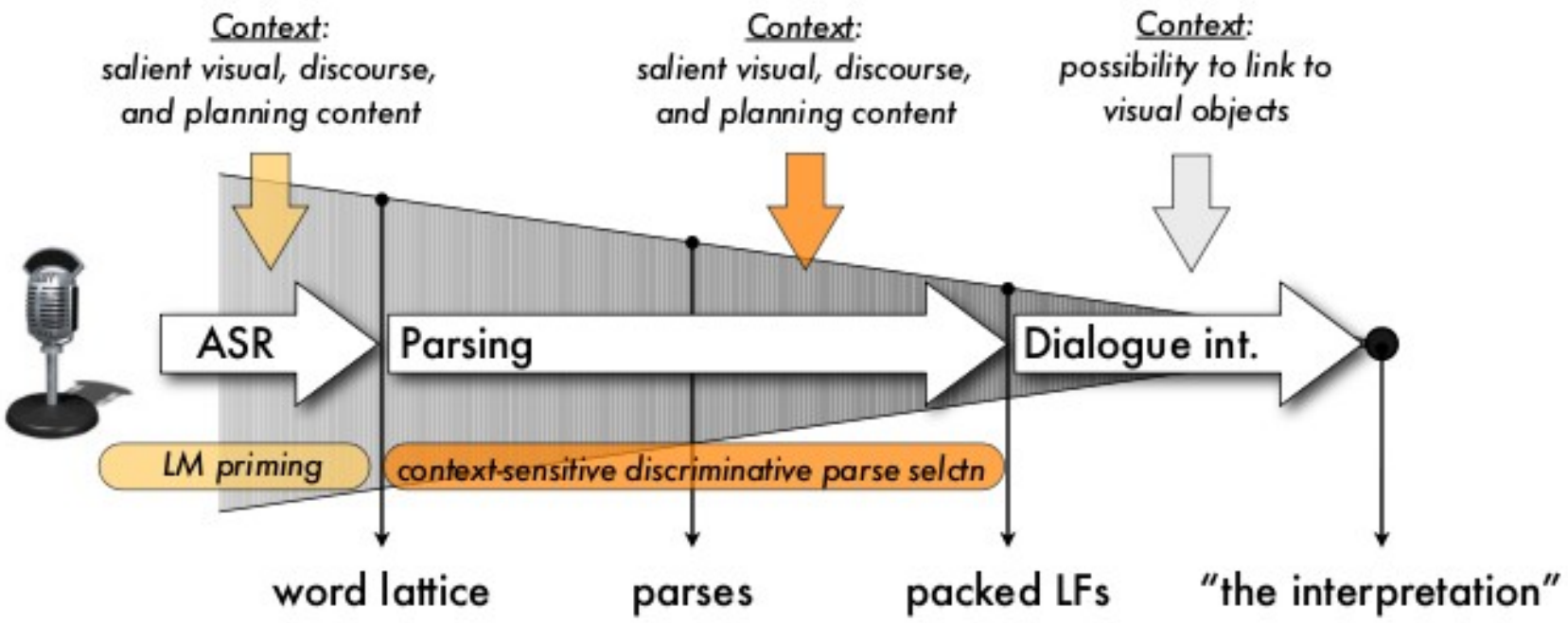
Overview of the approach



- Information about salient contextual entities (coming from the dialogue history or the immediate physical environment) are exploited to guide the speech recognition
- *Objective*: establish expectations about uttered words which are most likely to be heard given the context

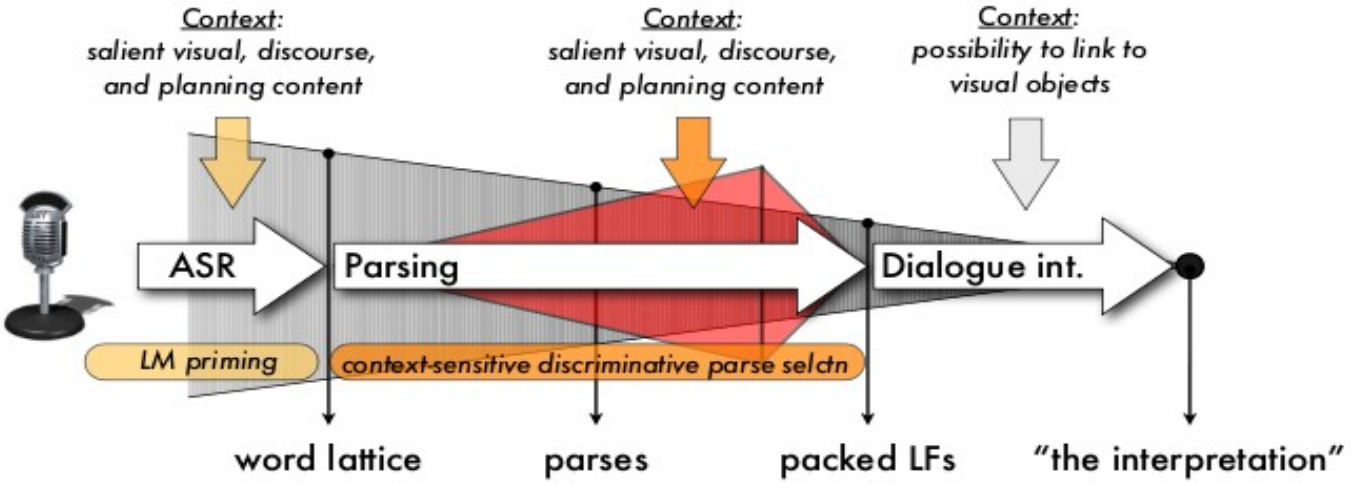
Overview of the approach





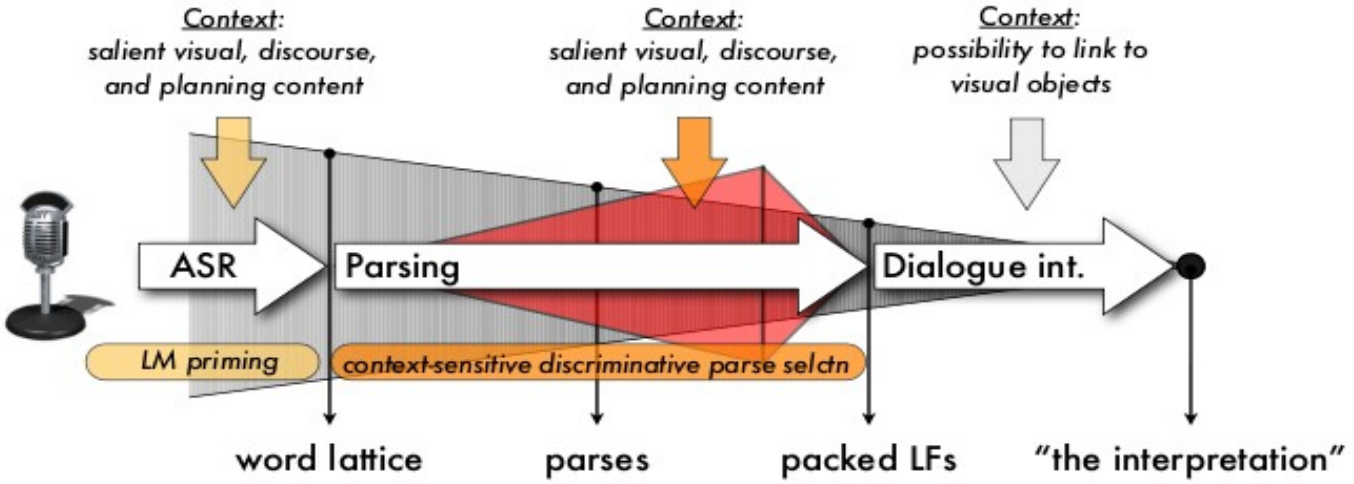
- Incremental parsing with Combinatory Categorial Grammar
  - Grammar able to handle ill-formed and misrecognised utterances by selectively relaxing and extending its set of grammatical rules.
- Use a discriminative parse selection model to select the most likely parse(s)
  - which includes various contextual features to guide the selection

Overview of the approach



- Incremental integration of context, ASR and parsing
  - Contextual information is used to balance the word lattice, priming what sequences are most likely in the considered context
  - Incremental parsing processes the word lattice
  - The result is a packed representation of possible interpretations
  - … which is then incrementally filtered by the parse selection module to retain only the most likely partial interpretations

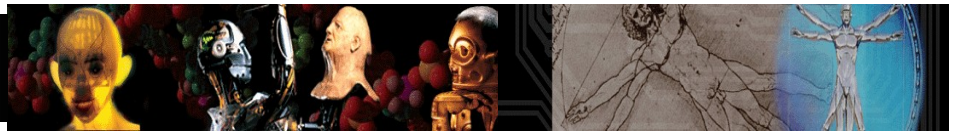
Overview of the approach

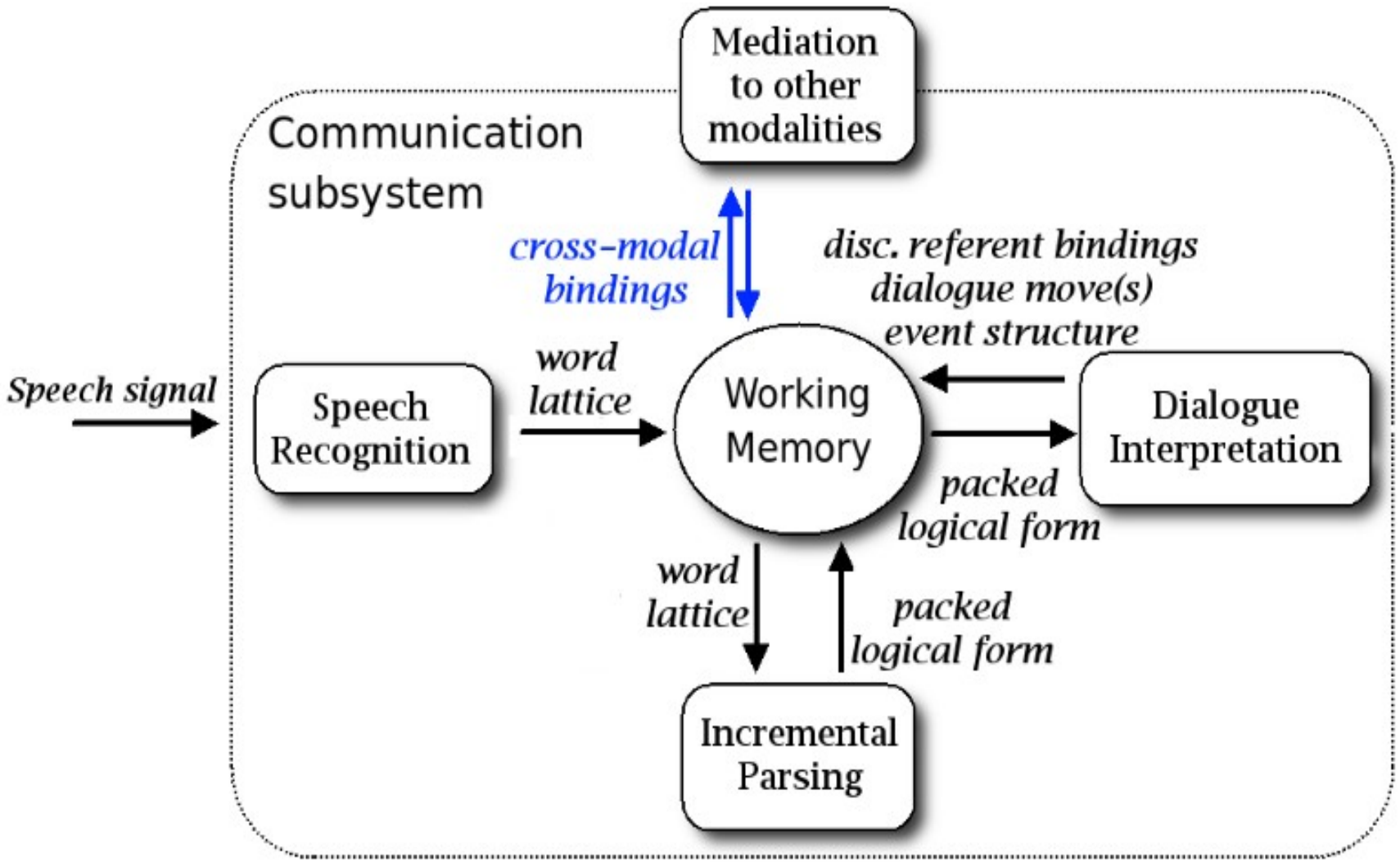


In three keywords:

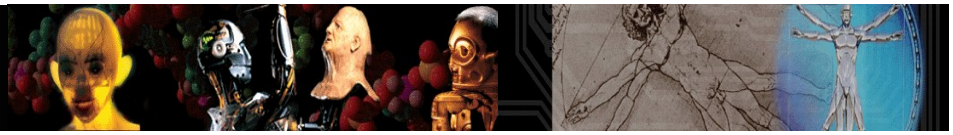
- **Hybrid**: Combination of fined-grained linguistic resources with statistical models, able to deliver both *deep* and *robust* dialogue processing
- **Integrated**: goes all the way from the speech signal up to the semantic and pragmatic interpretation
- **Context-sensitive**: Context is used at every processing step to guide the comprehension, both an *anticipation* tool and a *discrimination* tool

From an architectural perspective...



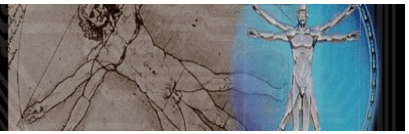


Outline of the talk

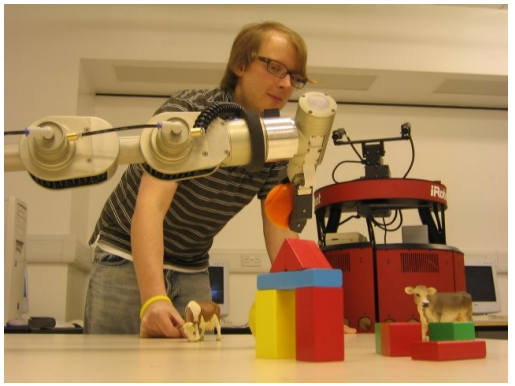


- What is human-robot interaction?
- Overview of the approach
- **Implementation**
  - Situated speech recognition
  - Robust parsing of spoken dialogue
  - Improving efficiency with chart pruning
- Conclusion

Context-sensitive ASR



- Priming speech recognition by exploiting context
  - Visuo-spatial context: objects in the visual scene
  - Dialogue state: expressions in the dialogue model
  - Objects are ranked according to saliency, and integrated into a cross-modal salience model
  - The cross-modal salience model primes lexical activations, which are incorporated into the language model for speech recognition.
- Lexical activation
  - A lexical activation network lists for each entity, the set of words activated by it, i.e. “those most likely to be heard”
  - This can include words related to object denomination, subparts, common properties, affordances.
- The contextual model is dynamically updated as the environment evolves
- Evaluation: **-16.1%** reduction in WER over baseline (Nuance ASR) [Lison 2008]



EXAMPLE

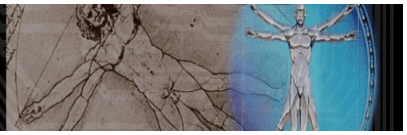
• Given a visually salient red block on a table top

• Recognized by the robot's visual system as such

• Lexical associations connect this “red block” to words like “block”, “square”, “pick up”, etc.

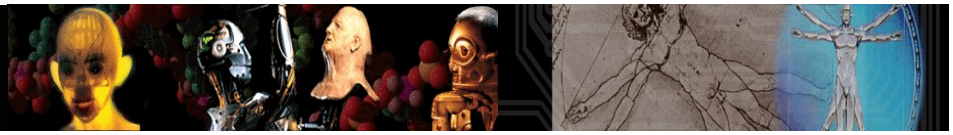
• The language model of the ASR is adapted to increase the probability of hearing these words

Robust parsing of spoken dialogue



- Difficulty of parsing spoken input
  - Parsing needs to be robust to *ill-formed* and *misrecognized* input
  - Different approaches possible: Shallow parsing, statistical approaches, controlled relaxation of grammar rules
  - **Grammar relaxation** through non-standard CCG rules added in the grammar; inspired by Zettlemoyer & Collins
- Different types of rules:
  - *Type-shifting rules* to account for missing words
  - *"Paradigmatic heap"* rules for dealing with syntactic disfluencies
  - *Discourse-level composition rules* for combining discourse units
  - *ASR correction rules* for correcting misrecognized words
- Problem: better coverage and integration, but also more analyses

Discriminative parse selection

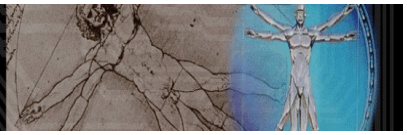


- Parsing produces a large number of analyses, arising from word lattice (multiple recognition hypotheses), controlled relaxation, and inherent ambiguity
- Need a mechanism to *filter/select* the resulting interpretations
- The task is represented as a function $F : X \rightarrow Y$ where the domain X is the set of possible inputs (word lattices), and Y the set of parses.
- The function F , mapping a word lattice to its most likely parse, is then defined as :

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

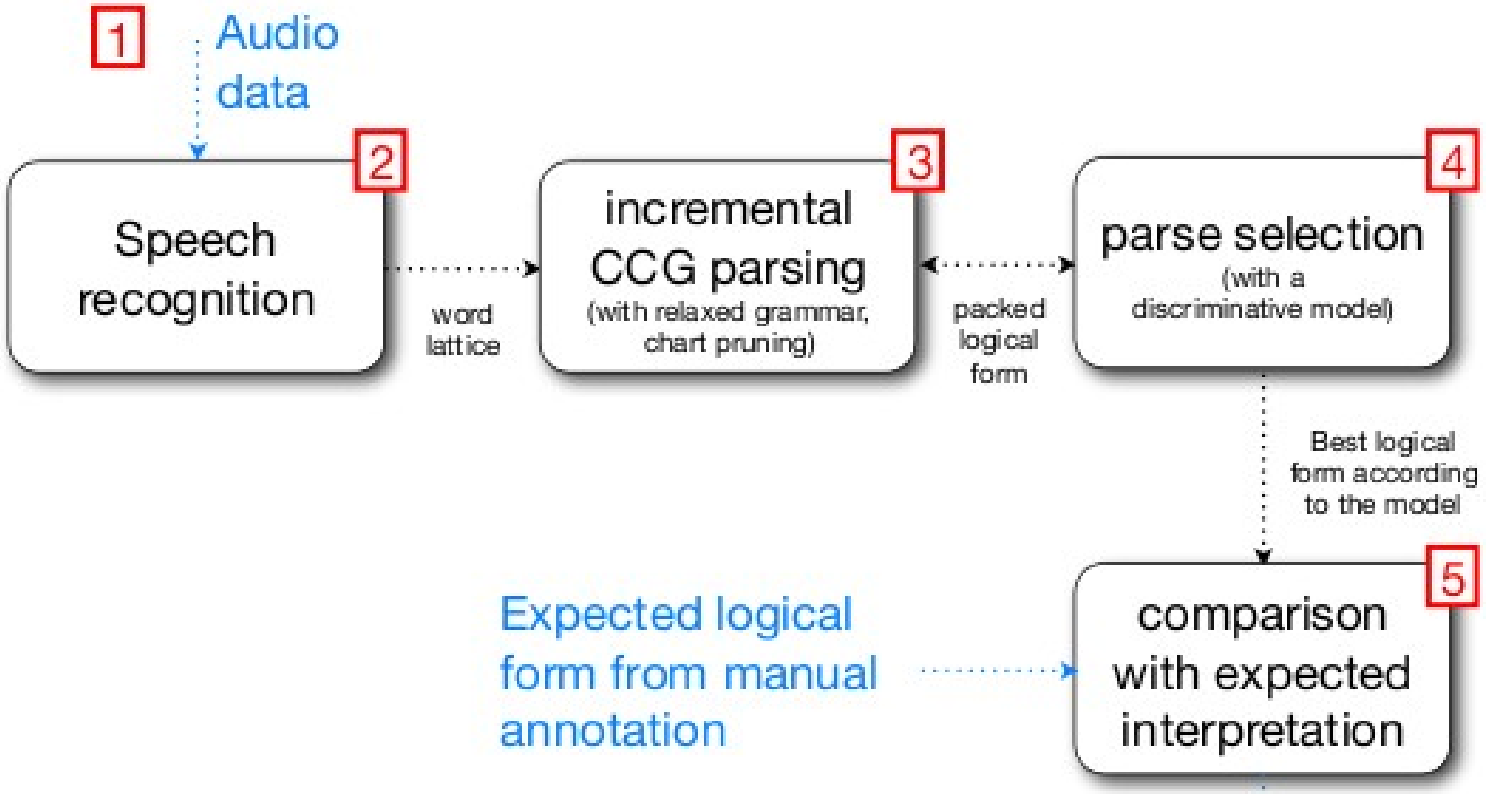
where $\mathbf{w}^T \cdot \mathbf{f}(x, y)$ is an inner product which can be seen as a measure of the "quality" of the parse.

Discriminative parse selection (2)



- Given the parameters **w**, the optimal parse of a given word lattice *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.
- Features include:
  - *acoustic features*: scores from speech recognition
  - *syntactic features*: derivational history of the parse
  - *semantic features*: substructures of the logical form
  - *contextual features*: situated and dialogue contexts
- The parameter vector **w** is learnt using a simple online perceptron
- Training on a corpus of automatically generated samples using a small domain-specific grammar

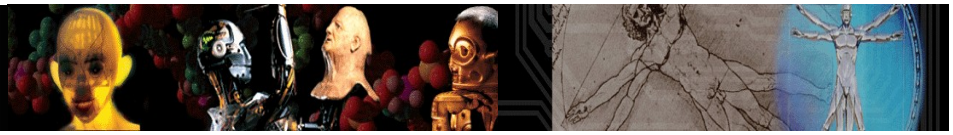
Evaluation setup



WoZ corpus of human-robot interactions:

- Tabletop scenario: visual object learning, manipulation
- Participants were free in both content and form - they could include questions, assertions, commands, answers, etc.
- Test suite was 195 utterances, manually annotated

Evaluation results



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

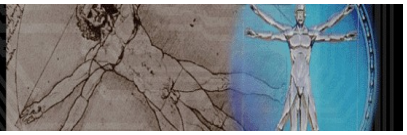
Exact-match accuracy results in %
(NBest 5 with all features activated)

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

Partial-match accuracy results in %
(NBest 5 with all features activated)

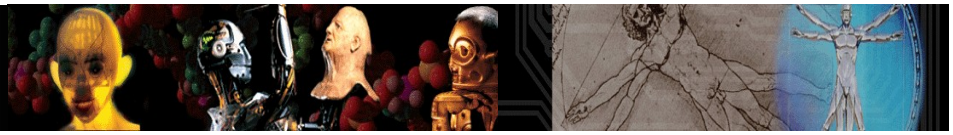
- + significant decrease of the word error rate, going from **20.5** % for the baseline to **15.7** % with our approach

Improving parsing efficiency



- Using parse selection *during* incremental parsing
- Rank partial analyses during parsing and perform incremental chart pruning
  - Fixed beam contains the highest ranked analyses (width ≈ 30)
  - Analyses outside the beam are pruned from the chart
- Effects of chart pruning
  - Chart pruning bounds parsing time and space complexity (which is crucial for real-time dialogue processing in HRI)
  - Evaluations show statistically significant reductions in parsing time, with virtually no loss in accuracy
- Empirical results on a WoZ corpus demonstrate a **53.8%** decrease in parsing time!

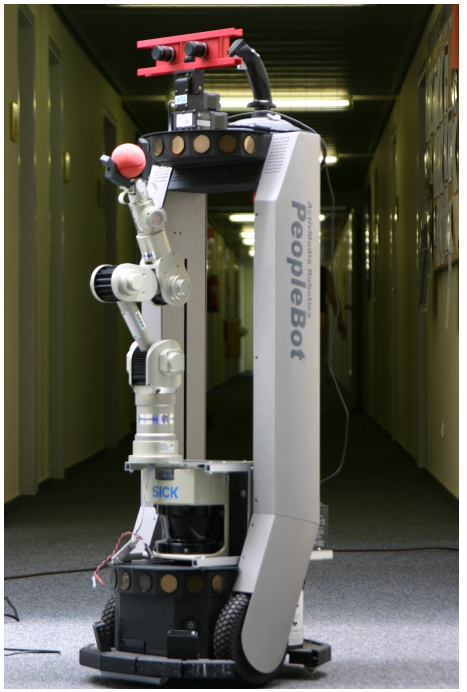
Outline of the talk



- What is human-robot interaction?
- Overview of the approach
- Implementation
  - Situated speech recognition
  - Robust parsing of spoken dialogue
  - Improving efficiency with chart pruning
- **Conclusion**

Conclusions and open issues

- Incrementality in parsing spoken dialogue
  - Incrementality allows for “context” to prime comprehension during processing
  - Priming yields significant improvements on WER, parsing performance, and parsing speed
  - Priming based on: context-sensitive language models, context-sensitive parse selection models, chart pruning
- Open issues
  - ASR not (yet) fully incremental: what if? How to deal with changing hypotheses - re-analysis, delaying analysis?
  - “Grounding” the selected interpretation in context, and using the resulting anchoring to further refine
  - (More in general: modeling conditions on referentiality, presupposition -- the intentional dimension of meaning)



For more information, check our website:

http://talkingrobots.dfki.de