Robust Processing of Spoken Situated Dialogue

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

Pierre Lison
Web: http://www.dfki.de/~plison

CogX Project
Cognitive Systems that Self-Understand and Self-Extend
EU FP7 IST Integrated project

Background

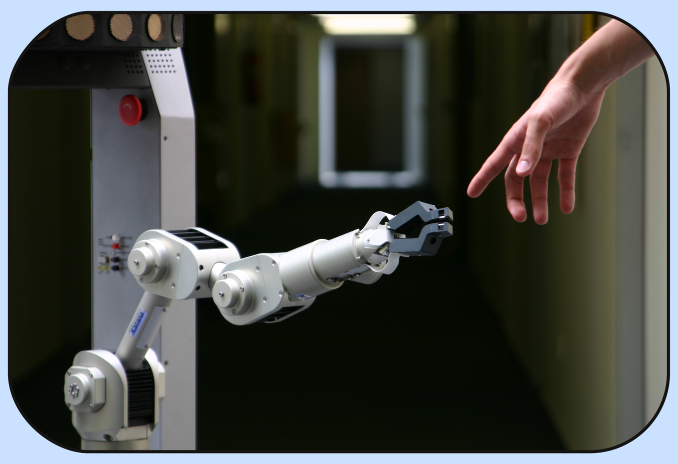
Human-Robot Interaction (HRI)

- HRI = research field dedicated to understanding, designing, and evaluating robotic systems for use by and with humans

- Multidisciplinar field: artificial intelligence, robotics, natural language processing, cognitive science, psychology

- We focus on one particular communication medium between a robot and a human: spoken dialogue

- How can we develop robots capable of understanding (and producing) situated, spoken dialogue?



Dialogue systems for HRI

- Spoken dialogue is a very natural mean of interaction between a robot and a human, but is difficult to process automatically

- Understanding the speech chain is not enough: the robot needs to relate the dialogue to an active understanding of its physical and social environment (what is the world around me, what can/should be done in this context, etc.)

- Dialogue systems for HRI must therefore be part of a larger cognitive system integrating perception, reasoning, and action

Key idea

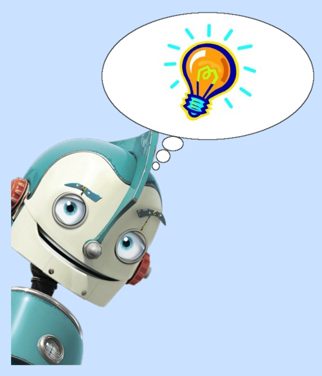
The Issue

Processing spoken dialogue is a challenging task:

1) Spoken utterances are often noisy, fragmentary, ambiguous, ungrammatical, and replete with disfluencies (filled pauses, speech repairs, repetitions, corrections)

2) Pervasiveness of speech recognition errors (word error rate typically in the 10-30 % range for non trivial domains)

→ Spoken dialogue systems must therefore be robust to both ill-formed and ill-recognised inputs



Our Solution

- Use a robust incremental parser able to handle ill-formed and misrecognised utterances by selectively relaxing its set of grammatical rules. The parser takes word lattices as inputs and generates a set of partial semantic interpretations

- The choice of the most relevant interpretation is then realised via a (statistical) discriminative model coupled to the parser. The discriminative model incorporates a broad range of linguistic and contextual features.

Approach

Architecture

- Our approach is implemented as part of a distributed, cognitive architecture encompassing several cooperating subsystems for communication, vision, motor control, and deliberative reasoning

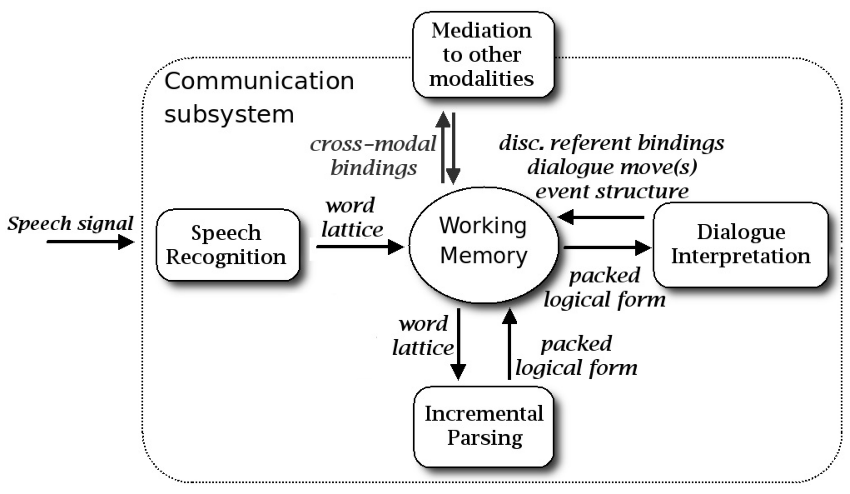
- The comprehension of a spoken utterance proceeds as follows:

STEP 1: the speech recogniser processes the audio signal to establish a word lattice containing ranked hypotheses about word sequences

STEP 2: A set of syntactic and semantic analyses (specified in a packed logical form) are constructed for the word lattice, using an incremental chart parser for Combinatory Categorial Grammar

STEP 3: the logical forms are resolved against a dialogue model to establish co-references and interpret dialogue moves

STEP 4: the linguistic interpretations are associated with extra-linguistic knowledge via a cross-modal information binding module



Architecture schema of the communication subsystem (limited to comprehension)

Grammar relaxation

- The grammatical constraints specified in the CCG grammar can be relaxed to handle slightly ill-formed or misrecognised utterances.

- Practically, the relaxation is realised via the introduction of non-standard rules in the CCG grammar (Zettlemoyer & Collins 2007). The rules can be grouped in three families:

> discourse level composition rules (to combine discourse units)
> "paradigmatic heap" rules (to handle disfluencies)
> ASR correction rules (to correct ASR errors)

Parse selection

- The grammar relaxation leads to a larger number of parses → we need a mechanism to discriminate the resulting interpretations

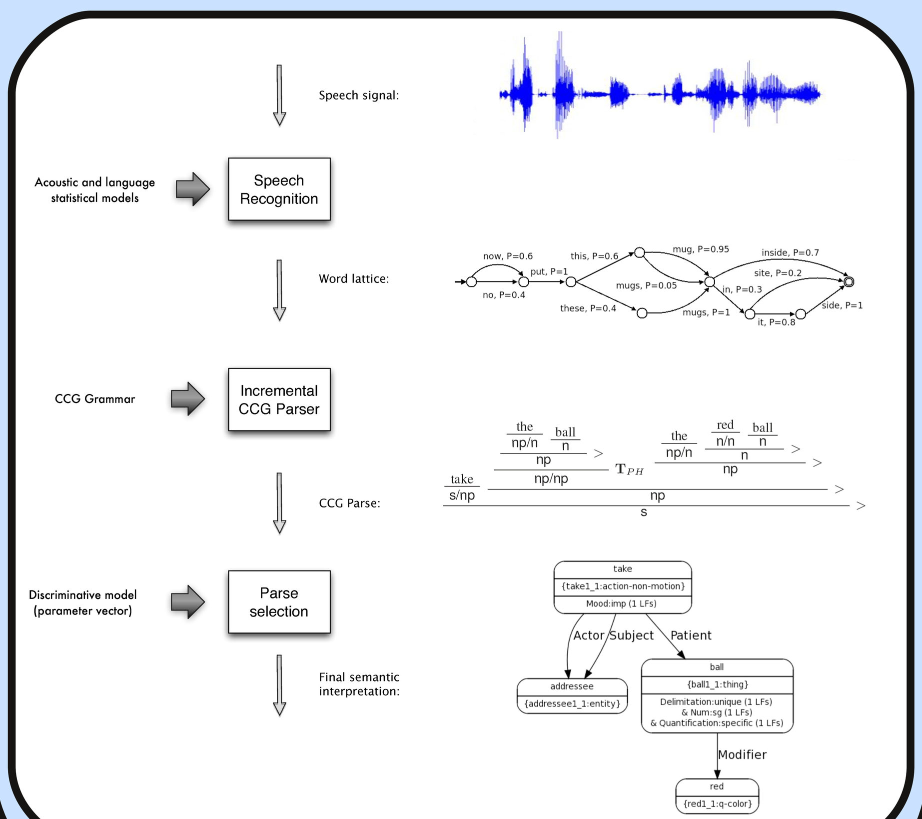
- Formally: a function F mapping a word lattice x to its most likely parse:

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

where: GEN(x) enumerates all possible parses for x
$f(x,y) \in \Re^d$ is a vector representing features of the pair (x,y)
$w \in \Re^d$ is a parameter vector

- The feature fector f(x,y) includes:
> Semantic features (substructures of the logical form)
> Syntactic features (derivational history of the parse)
> Acoustic features (speech recognition scores)
> Contextual features (situated and dialogue context)

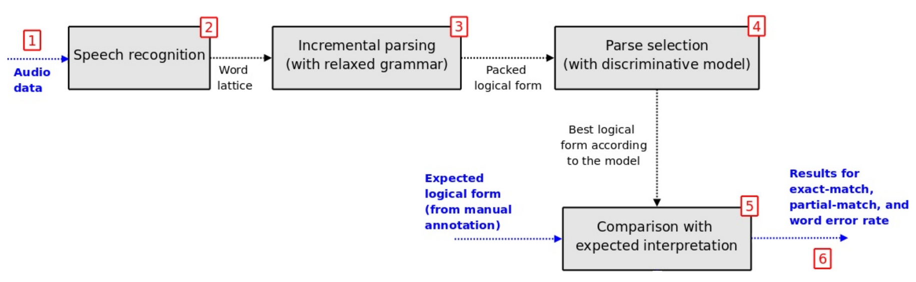
Example



Evaluation

Experiment Setup

- We performed a quantitative evaluation of our approach based on a collected Wizard-of-Oz corpus oh human-robot spoken dialogue (195 utterances manually segmented and annotated) for a task domain of object manipulation and visual learning



- Three types of results are extracted: exact-match, partial match, and word error rate

Results

	Size of word lattice (number of NBests)	Grammar relaxation	Parse selection	Precision	Recall	F_1-value
(Baseline)	1	No	No	40.9	45.2	**43.0**
.	1	No	Yes	59.0	54.3	56.6
.	1	Yes	Yes	52.7	70.8	60.4
.	3	Yes	Yes	55.3	82.9	66.3
.	5	Yes	Yes	55.6	84.0	66.9
(Full approach)	10	Yes	Yes	55.6	84.9	**67.2**

Table 2. Exact-match accuracy results (in percents).

	Size of word lattice (number of NBests)	Grammar relaxation	Parse selection	Precision	Recall	F_1-value
(Baseline)	1	No	No	86.2	56.2	**68.0**
.	1	No	Yes	87.4	56.6	68.7
.	1	Yes	Yes	88.1	76.2	81.7
.	3	Yes	Yes	87.6	85.2	86.4
.	5	Yes	Yes	87.6	86.0	86.8
(Full approach)	10	Yes	Yes	87.7	87.0	**87.3**

Table 3. Partial-match accuracy results (in percents).

Size of word lattice (NBests)	Grammar relaxation	Parse selection	WER
1	No	No	**20.5**
1	Yes	Yes	19.4
3	Yes	Yes	16.5
5	Yes	Yes	15.7
10	Yes	Yes	**15.7**

Table 4. Word error rate (in percents).

statistically significant improvements both in robustness and accuracy over the baseline