EU FP7 CogX
ICT-215181
May 1 2008 (52months)

DR 6.1: Transparency in situated dialogue for interactive learning (in human-robot interaction)

Geert-Jan M. Kruijff, Miroslav Janiček, Ivana Kruijff-Korbayová, Pierre Lison, Raveesh Meena, Hendrik Zender

DFKI GmbH, Saarbrücken
⟨gj@dfki.de⟩

Due date of deliverable:	July 31 2009
Actual submission date:	July 24 2009
Lead partner:	DFKI
Revision:	final
Dissemination level:	PU

A robot can use dialogue to try to learn more about the world. For this to work, the robot and a human need to establish a mutually agreed-upon understanding of what is being talked about, and why. Thereby it is particularly important for the human to understand what the robot is after. The notion of *transparency* tries to capture this. It involves the relation between why a question is asked, how it relates to private and shared beliefs, and how it reveals what the robot does or does not know. For year 1, WP6 investigated means for establishing transparency in situated dialogue for interactive learning. This covered two aspects: how to phrase questions for knowledge gathering and -refinement, and how to verbalize knowledge. Results include methods for verbalizing what the robot does and does not know about referents and aspects of the environment, based on a mixture of prior and autonomously acquired knowledge and basic methods for self-understanding (Task 6.1); and, novel algorithms for determining content and context for question subdialogues to gather more information to help resolve misunderstandings or fill gaps (Task 6.2). WP6 also reports results on making spoken situated dialogue more robust, employing probabilistic models for using multi-modal information to reduce uncertainty in comprehension.

Executive Summary

One of the objectives of CogX is self-extension. This requires the robot to be able to actively gather information it can use to learn about the world. One of the sources of such information is dialogue. But for this to work, the robot needs to be able to establish with a human some form of mutually agreed-upon understanding – they need to reach a *common ground*. The overall goal of WP6 is to develop adaptive mechanisms for situated dialogue processing, to enable a robot to establish such common ground in situated dialogue.

WP6 primarily focuses on situated dialogue for continuous learning. In continuous learning, the robot is ultimately driven by its own curiosity, rather than by extrinsic motivations. The robot builds up its own understanding of the world – its own categorizations and structures, and the ways in which it sees these instantiated in the world. While learning, the robot can solicit help from the human, to clarify, explain, or perform something. This is where situated dialogue can help the robot to self-extend – and which is where transparency comes into play. The robot is acting on its own understanding, which need not be in any way similar to how a human sees the world. There is therefore a need for the robot to make clear what it is after: why the robot is requesting something from a human, what aspects of a common ground it appeals to, and how the request is related to what it does and does not know.

To achieve transparency in situated dialogue for continuous learning, WP6 investigated two important aspects in year 1: Verbalization of knowledge about classes and instances (Task 6.1), and phrasing questions as subdialogues (Task 6.2). WP6 developed novel methods for context- and content-determination in verbalizing knowledge about referents and aspects of the environment, with the possibility to combine a priori and autonomously acquired knowledge. As a result, the robot is capable to refer to instances in a contextually appropriate way, phrase their description relative to knowledge about other instances and classes, and talk about ontological knowledge it has. We base some of these methods in simple ways for the robot to introspect what it knows about an entity (self-understanding), and establish gaps in its understanding of that entity relative to ontological categories. Connected to these efforts, WP6 developed new algorithms for context- and content-determination for question subdialogues, setting such determination against the background of context models of multi-agent beliefs and intentions; and for realizing these dialogues with contextually appropriate intonation. The robot can now plan for how to request information from the user to clarify or extend its understanding. It does so in a manner that appropriately reflects how this request relates to private and shared beliefs, and intentions. In such a mixed-initiative dialogue, the robot can dynamically adapt its plan to achieve its knowledge-gathering intention.

In addition to the main focus on transparency, WP6 continued efforts in making spoken situated dialogue more robust. Further improvements in robustness were achieved by using context information in incremental, distributed processing covering speech recognition, parsing, and dialogue interpretation. This approach explicitly deals with the uncertainty in the speech recogniser in a Bayesian way, which is part of our broader approach in CogX to using probabilistic representations to capture uncertainty in initial interpretations of sensory signals. By maintaining all the possible hypotheses we can then use knowledge from other modalities to revise our interpretation or to bias inference. This is an example of a simple kind of self-understanding, since we are representing the possibilities and uncertainties in our interpretations.

Role of (transparent) situated dialogue in CogX

CogX investigates cognitive systems that self-understand and self-extend. In some of the scenarios explored within CogX such self-extension is done in a mixed-initiative, interactive fashion (e.g. the George and Dora scenarios). The robot interacts with a human, to learn more about the environment. WP6 contributes situated dialogue-based mechanisms to facilitate such interactive learning. Furthermore, WP6 explores several issues around the problems of self-understanding and self-extension in the context of dialogue processing. Dialogue comprehension and production is ultimately based in a belief model the robot builds up. This belief model captures beliefs and tasks, in a multi-agent fashion. We can attribute a belief/task to one agent (private), multiple agents (shared), or have an agent attribute a belief/task to another agent (private, attributed). Already at this level we thus see a range of possible forms of self-understanding and self-extension. The goal of transparency is to establish beliefs as shared, and thus, any belief that should be shared but currently is not represents a gap of sorts. The differentiation between private and shared status status is one aspect of context that helps determine how we produce references to entities in the world, and the way we produce questions about such entities. Furthermore, interpretations leading up to these beliefs and tasks may be uncertain. We use probabilistic models to help counter uncertainty in comprehension, fusing information from multiple modalities to guide comprehension. Should this fail, we can use clarification to overcome that uncertainty. Such clarification can also be used to resolve uncertainty about situated understanding, or in a more general way, to request information about entities in the world. WP6 presents a first attempt at an algorithm for identifying gaps in terms of unknown properties about an entity I relative to a category C. We use these gaps as a basis for verbalizing what a robot does and does not know about I, and to drive dialogue to gain more information about I.

Contribution to the CogX scenarios and prototypes

WP6 contributes directly to the George and Dora scenarios, in relation to work performed in WP 3 (Qualitative spatial cognition), WP 5 (Interactive continuous learning of cross-modal concepts), and WP 7 (Scenario-based integration). Robust dialogue processing, clarification, and verbalization are in principle used in both scenarios. In George we provide the possibility for the robot to ask about visual properties it is uncertain about, and to use verbalization and referencing to describe what it sees:

- **Human** places a red box on the table
- **Robot** Vision recognizes the object as a box, but is unsure about the color. A clarification request is triggered, handled by dialogue.
- **Robot** "Is that box red?" – dialogue provides indirect feedback it has recognized the object as a box, while at the same time asking for confirmation on the color.
- **Human** "Yes, this box is red."
- **Robot** Vision is provided with the information that the box is indeed red, and so can update its models.

In Dora we also explore the introspection on what the robot does and does not know about an area, to drive information requests to the user. (The method is in fact general enough to also drive active visual search in the environment.)

- **Human** guides the robot to a new area, and says "Here we are in the kitchen." This the second kitchen the human and the robot visit.
- **Robot** Place categorization can determine the area as a kitchen, with a particular size. Vision perceives a water cooker.
- **Robot** "Ok, this looks like a larger kitchen." – the robot can compare to other kitchen instances it has seen so far.
- **Robot** The robot can infer that kitchens typically have several objects, not only a water cooker but also a coffee machine. It understands that it does not know of a coffee machine here, though.
- **Robot** "I can see a water cooker. Is there also a coffee machine?" – the robot indicates what it does and does not know, and uses this as the background for extending its knowledge about the area.

1 Tasks, objectives, results

1.1 Planned work

Robots, like humans, do not always know or understand everything. Situated dialogue is a means for a robot to extend or refine its knowledge about the environment. For this to work, the robot needs to be able to establish with a human some form of mutually agreed-upon understanding – they need to reach a *common ground*. The overall goal of WP6 is to develop adaptive mechanisms for situated dialogue processing, to enable a robot to establish such common ground in situated dialogue.

WP6 primarily focuses on situated dialogue for continuous learning. In continuous learning, the robot is ultimately driven by its own curiosity, rather than by extrinsic motivations. The robot builds up its own understanding of the world – its own categorizations and structures, and the ways in which it sees these instantiated in the world. While learning, the robot can solicit help from the human, to clarify, explain, or perform something. This is where transparency comes into play. The robot is acting on its own understanding, which need not be in any way similar to how a human sees the world. There is therefore a need for the robot to make clear what it is after: why the robot is requesting something from a human, what aspects of a common ground it appeals to, and how the request is related to what it does and does not know. To achieve transparency in situated dialogue for continuous learning, WP6 investigated two important aspects in year 1: Verbalization of knowledge about classes and instances (Task 6.1), and phrasing questions as subdialogues (Task 6.2).

Task 6.1: Verbalising categorical knowledge *The goal is to enable the robot to verbalize its own categorical knowledge (or lack thereof) relative to a situation, and understand situated references. We will extend existing methods for comprehending and producing referring expressions to cover verbalization of relevant information from singular visual categories (WP5) and contextual reference.*

Task 6.2: Continual planning for clarification and explanation *We will extend strategies for planning clarification- and explanation dialogues using a continual planning approach. This offers the necessary flexibility to adjust a plan when interactively setting up an appropriate context, and provides a model of common ground in dialogue. These methods will be based in means for grounding the information expressed by clarifications and explanations in situated understanding.*

The intention behind Tasks 6.1 and 6.2 was to achieve that the robot would be able to enter into a dialogue with a human, to clarify something or to request more information. This could be either about dialogue itself, or regard the situated context being talked about – thus spanning the entire range of Clark's grounding levels [11]. The robot uses belief models to represent private and shared beliefs, including private beliefs the robot attributes to other agents, and ontologies to capture its categorical knowledge about the world. Together, belief models and ontologies provide a rich epistemological background against which the robot can introspect what it does or does not know (e.g. whether another agent does understand something, or whether an observed instance is of a particular category). We use such self-understanding to guide verbalization and clarification, two interrelated functions to help the robot gather more information to self-extend. The role of verbalization in this process is to ensure that the why what and how of the question is clear to the human: why the robot asks, what it does and does not know, and how that gap should be addressed. The planning part is to take care of the planning and execution of the actual dialogue, to ensure human and robot eventually do achieve a common ground. In §1.2 we describe how we achieved these goals.

1.2 Actual work performed

Below we succinctly describe the achievements for the individual tasks. The descriptions refer to the relevant papers and reports in the annexes, for more technical detail. In §1.3 we place these achievements in the context of the state-of-the-art.

1.2.1 Verbalising categorical knowledge

The goal of Task 6.1 was to develop methods for the robot to verbalize its own categorical knowledge, or lack thereof. We have achieved the following:

Context-determination, bi-directionality in referencing A robot typically acts in an environment larger than the immediately perceivable situation. The challenge in referring to objects and places in such a large environment is to ensure that the agents participating in the dialogue can identify the appropriate context against which the resolve a reference. Zender et al (§2.1.1, §2.1.2) have developed novel methods for determining the appropriate context for comprehending and producing referring expressions.

A typical example Zender et al address is when the robot needs to refer to an object in a place other then where the robot currently is, talking to a human. Or when it needs to understand such a reference. For example, the robot has been sent to fetch a person to take a phone call in somebody else's office (e.g. GJ's). If this person is currently in her office, it would not do to say "there's a call for you on the phone." This could incorrectly identify the phone on that person's desk as the one to pick up, whereas the point is to go to the GJ's office to take the call there. What Zender et al do is to use topological structure of the environment, to –literally– determine the appropriate context for identifying the object it needs to refer to. So, instead of just saying "the phone," the robot is able to say "there is a call for you on the phone on the desk in GJ's office." It uses the context to direct the human's attention to the appropriate location, where it can then identify the intended referent.

Verbalization of acquired properties Typically a robot is provided with an initial ontology, outlining the concepts and their relations considered relevant for understanding the environment in which the robot is to act. Over time, the robot can extend this ontology, for example with instances and properties that hold for these. Zender and Pronobis (§2.1.3) have developed a new method for verbalizing knowledge about autonomously acquired scalar properties for instances and their classes. The distributions of property values across instances, within a class, and across classes help define contextual standards [34, 15] against which the verbalization of scalar properties as comparatives can be determined in a contextually appropriate manner.

A scalar property is, simply, a property with values that are on a scale that makes them comparable. An example of a scalar property is size: A room can be smaller or larger than some other room, or of the same size. Scalars are typical material properties for the kinds of entities we want the robot to talk about. And, they are properties for which the robot can autonomously acquire quantitative models. The problem is, how to then talk about them. We cannot simply verbalize such a property at face value, e.g. as "the room is $14.67m^2$." Humans prefer more qualitative descriptions, like "large" or "smaller." Such qualitative descriptions are called *vague scalar predicates.* Their exact interpretation is left vague – that is to say, their exact interpretation is relative to a particular *contextual standard* which defines the scale along which comparisons are to be made. Zender and Pronobis propose a method to make it possible for the robot to introspect what variation it has perceived for a particular scalar property among instances of a class, or among classes as such. This form of self-understanding enables the robot to talk in a human-like, qualitative fashion about scalar properties, while at the same time (indirectly) indicating to the human what it considers as prototypical values (by comparison).

Verbalization of categorical knowledge Sometimes it is more important for the robot to make clear what it does *not* understand, than to say what it does know about. This helps the human to figure out what the robot might be after. Zender and Kruijff (§2.1.4) discuss a preliminary method for a robot to introspect the knowledge it has

> about an entity in the world. The method establishes what the robot does and does not know about that entity relative to one or more categories in a known ontology. The resulting identified "gaps" are those properties for the entity that the robot does not know about, but which it would need to know to establish the entity as an instance of a particular category. Zender and Kruijff subsequently discuss how the robot can then verbalize this self-understanding, in terms of what the category, the instance and its known properties, and the missing properties identified as gaps.

Zender and Kruijff consider a simple, but often occurring form of "gap": namely, when a robot is lacking property information about an object or an area to fully determine whether it is an instance of a particular category. Consider again the example given earlier. A human and a robot enter a new room, which the human indicates is a kitchen. The robot can categorize the place as such, and even sees a water cooker. However, based on the knowledge it has about kitchens, it would also expect a coffee machine to be there. Zender and Kruijff show how the robot can determine such a property of "having a coffee machine" as a gap in its knowledge about this area (as being a kitchen). To convey this self-understanding, Zender and Kruijff discuss how the robot can then verbalize this gap, together with a description of what it does know about the area-as-a-kitchen. "Ok, this looks like a larger kitchen. [...] I can see a water cooker. Is there also a coffee machine?"

The novelty in all these methods is the role context plays in determining how a robot should understand or verbalize a reference, or what it knows about something (be that an instance or a class). Traditional methods focus primarily on *content*-determination, typically assuming a context to be given. Our methods combine content-determination with context-determination. Context-determination can thereby mean both situated context (e.g. references in large-scale space) and epistemological context (e.g. what beliefs a robot has, or attributes to other agents, or what it knows about how to compare across classes). With that we go beyond the original objectives of Task 6.1, which focused only on verbalizing knowledge about visual objects in a current scene.

1.2.2 Clarification

The goal of Task 6.2 was to develop methods so a robot could clarify or expand what it understands about the environment. These methods were to be continual, in the sense that it should be possible to monitor the execution of a plan, and where necessary adapt or expand it. We have achieved this goal in the following ways.

Determining epistemological context in questions Transparency and scaffolding in dialogue for learning depend on epistemological context: how questions appeal to common ground, what private beliefs they are based in – and what answers an interlocutor would like to have. Kruijff & Brenner (§2.2.1) explore methods for determining such appropriate epistemological contexts, considering transparency and scaffolding explicitly as referential qualities of a question. These contexts are then connected in a notion of question nucleus which reflects what is being asked after, in reference to beliefs (aboutness, transparency) and intentions (resolvedness, scaffolding). A question nucleus provides the basis for formulating a question as a dialogue.

A robot should not just go and blurt out a question – this may not lead to the human given the desired answer. A nice example of this is provided by the former CoSy Explorer system [37]. The robot classified every narrow passage it went through ($< 70cm$) as a door. Sometimes it would realize that some previous passage probably wasn't a door, just an artifact of driving around in a cluttered environment. At that point of realization, the robot would just ask "Is there a door here?" Out of the blue, without further indication of where there ought to be a door, a human would typically say "yes" – understanding the robot to mean, whether there would be a door to this room. Which, of course, was not what the robot meant. But what it failed to do was to properly take into account what the human would know (she didn't know that "here" was supposed to refer to that narrow passage), and how to formulate its question accordingly. Kruijff and Brenner look into how the robot could use its multi-agent belief models to determine how to best pose a question. They start by formulating ways for the robot to introspect its beliefs to determine what the human knows about something the robot wants to ask a question about. This determines how to refer the entity under discussion – making it transparent what the robot is talking about. A second step is to use what the robot holds as private knowledge and beliefs about the entity, to properly indicate what it would like to know more about.

Continual comprehension and production of clarification Brenner et al (§2.2.2) consider how a continual approach for planning and executing dialogues can be applied to human-robot interaction, in general. Kruijff & Janiček (§2.2.3) combine these insights with weighted abduction. The approach covers comprehending and producing dialogue and combines intention, attentional state, and multi-agent belief modeling. Kruijff & Janiček focus on clarification dialogues, covering Clark-style grounding from communicative levels to information requests concerning situated understanding.

Robots don't always understand everything. Sometimes they realize that, but sometimes they don't, and attribute some property to an object that is just plain wrong. Kruijff & Janiček try to capture such forms of collaboration between a human and a robot, dealing explicitly with the continual nature of such collaboration – things may go wrong and then need to be corrected. A typical example that they try to capture is the following:

(1) Human places an object on the table

(2) Robot: "That is a brown object."

(3) Human: "It is a red object."

(4) Robot: "Ok. What kind of object is it?"

(5) Human: "Yes."

(6) Robot: "Aha. But what KIND of object is it?"

(7) Human: "It is a box."

Kruijff & Janiček explicitly use the belief models of the robot, for the robot to figure out how it could use beliefs and observations to establish why a human may have said something, and how to best achieve what the robot itself is after (in terms of updating its beliefs). They make it possible for the robot to assert a belief ("this is a brown object") but then having to retract it when being corrected by a human ("it is a red object") and establishing the corrected belief as a shared belief about the scene ("ok"). At the same time, using what it understands to be shared, the robot can make safe assumptions about how it can refer to objects. Attributed beliefs also make it possible for the robot to assume that the human may know an answer to a question. In its reasoning the robot can then assert that the human will provide it with that information ("what kind of object is it?"). With that the robot first of all explicitly represents the gap in its knowledge (what it would like to know). But this also provides a level at which introspection can track the extent to which the gap has actually been resolved. The robot checks the updates it can make to its belief model in response to its question, and can use the "self"-understood failure to do so to persist in trying to get an appropriate answer from the human. Humans are not always fully cooperative, so when the human replies with "yes" (as in "coffee or tea? yes please") she does not provide an answer to the question. (Non-cooperative behavior is a problem usually "assumed away" in approaches to dialogue; Kruijff & Janiček don't, dealing with it in a continual way as argued for in Brenner et al, §2.2.2.) The robot can figure this out (using the approach of Kruijff & Brenner, §2.2.1), repeat the question, to then finally get the desired kind of answer ("it is a box.").

Contextually appropriate intonation for questions Kruijff-Korbayová et al (§2.2.4) develop new methods for determining information struc-

> ture and intonation for a range of utterance types, including commands, assertions, and -most importantly- questions. Information structure and its reflection through e.g. intonation can make it clear to the hearer how the utterance relates to the preceding context, and what it focuses on. As with assertions, where intonation can change the dynamic potential of their interpretation, intonation in a question indicates its dynamics in terms of what it is after: what type of answers is expected. Kruijff-Korbayová et al outline experiments to verify these theoretical insights in an empirical way.

There is more to saying something than simply uttering a sequence of words. In English, the intonation of an utterance reflects what it is that someone is talking about, and what she would like to focus on. There is a marked difference between assertions like “this is a RED box” (capitalization indicating stress) versus “this is a red BOX,” or questions like “is this a red BOX?” or “is this is a RED box?” Getting this right is crucial for the robot to convey what it is after. Kruijff-Korbayová et al (§2.2.4) describe how the robot can use private and shared beliefs, and what is currently attended to, to help drive how to formulate a contextually appropriate intonation for an utterance. In combination with the previous achievements, this rounds it all off: We can determine what beliefs and gaps play a role in formulating e.g. a question, we can manage a dialogue around that question, we can verbalize its content and references in a contextually appropriate way, and formulate all that with the right intonation.

The novelty in all these methods is thus how they achieve to flexibly combine intention, multi-agent beliefs and attentional state in continual processing of dialogue. Based on existing approaches, these methods explore how the robot can introspect the private and shared beliefs it entertains, situate beliefs and intentions, and then use that as a background against which it can handle and overcome pervasive aspects such as uncertainty, and the typically large-scale spatiotemporal nature of action and interaction.

1.2.3 Robust processing of spoken situated dialogue

The success of dialogue-based human-robot interaction ultimately stands or falls with how well a robot understands what a human says. Unfortunately, spoken dialogue is difficult to understand. Utterances are typically incomplete or contain disfluencies, they may be ungrammatical, or a speaker may correct herself and restart part of an utterance. This requires processing of spoken dialogue to be robust. At the same time, we cannot sacrifice deep understanding for robustness, as is often done. In the end a robot needs to understand what a human said, to be able to act on it. That is the whole point of situated dialogue as we consider it here.

In addition to Tasks 6.1 and 6.2, we have continued our efforts in robust processing of spoken situated dialogue. These efforts started already in CoSy; the results reported here build up on these previous efforts but have been achieved entirely during year 1 in CogX.

Integration of context in speech recognition and incremental parsing Lison & Kruijff (§2.3.1) present a novel approach in which context information is used in a process combining speech recognition and incremental parsing. The approach considers the entire processing from incoming audio signal to establishing a contextually appropriate interpretation of an utterance. Lison & Kruijff show that substantial improvements on robustness in processing can be achieved (measured against a WoZ corpus) by including context information (e.g. salient objects, actions). This information is used to bias lexical activation probabilities in the language model for speech recognition, and to guide discriminative models for parse ranking applied at the end of an incremental parsing process.

Incremental contextual pruning in parsing Lison & Kruijff (§2.3.2) consider the application of discriminative models during incremental parsing. After each step, context-sensitive discriminative models are applied to rank analyses. Using a beam of width 30, Lison & Kruijff show how parsing time can be reduced by 50% without suffering any significant reduction in performance (measured on a WoZ corpus).

When a human processes visually situated dialogue, she uses what she sees in the scene and how she knows that scene to help her understand what someone else might be saying about that scene. Lison & Kruijff explore how this idea can be used to make spoken dialogue processing in human-robot interaction more robust. When a robot perceives objects in a scene, it uses that information to activate expressions it could associate with such objects. For example, if it sees a ball, it would activate expressions like "round," "pick up," etcetera. These expressions are phrases the robot expects to hear. They help the robot to anticipate what a human is likely to say, when talking about that scene. Lison & Kruijff show that the robot can use this information to deal with the uncertainty inherent to speech recognition. Doing it in a probabilistic way, it is part of the broader approach in CogX to using probabilistic representations to capture uncertainty in initial interpretations of sensory signals. By maintaining all the possible hypotheses we can then use knowledge from other modalities to bias how the audio signal is interpreted in terms of possible word sequences. This is an example of a very simple kind of self-understanding, since we are representing the possibilities and uncertainties in our interpretations. Lison & Kruijff take this even further, by using the same information about the context to then help parsing to discriminate between possible analyses, to end up with a parse that represents the most likely semantic interpretation of the audio signal in the given context. Using this sort of discrimination-based-on-context during parsing actually helps to reduce the time needed to parse an utterance.

1.3 Relation to the state-of-the-art

Below we briefly discuss how the obtained results relate to the current state-of-the-art. We refer the reader to the annexes for more in-depth discussions.

1.3.1 Verbalisation

Task 6.1 considered the use of methods for comprehending and producing *referring expressions* to cover verbalization of knowledge, and contextual reference. For such expressions to appropriately refer to the intended referent, they need to meet a number of constraints, to help a hearer identify what is being talked about. First, an expression needs to make use of concepts that can be understood by the hearer. This becomes an important consideration when we are dealing with a robot which acquires its own models of the environment and is to talk about the contents of these. Second, the expression needs to contain enough information so that the hearer can distinguish the intended referent from other entities in the world or a belief state, the so-called *potential distractors*. For this it is necessary that the robot takes the differentiation between private and shared beliefs into account, as we already saw earlier. Finally, this needs to be balanced against the third constraint: Inclusion of unnecessary information should be avoided so as not to elicit false implications on the part of the hearer.

Zender & Pronobis (§2.1.3) particularly deal with the first aspect. Given that a robot autonomously acquires knowledge about the world, how can such properties be used to verbalize what the robot knows? Existing work on modeling *scalar properties* considers the use of *contextual standards*, to determine how to realize such properties as "vague" expressions involving gradable adjectives [15, 34]. This research primarily focuses on instances – in a given visual setting. Zender & Pronobis move beyond this, by considering how scalar properties can be modeled as probabilistic distributions over their values – and then use these distributions to construct contextual standards. This makes it possible to consider distributions solely across observed instances (like [15]), and also across instances within a class (considering values to be *prototypical* values within a class), and across classes. Within-class and across-class contextual standards are not considered (nor immediately possible) in [15]. They are, however, necessary to generate contextually appropriate verbalizations using comparatives. For example, consider the average office to have $8m^2$. Talking about two offices, with *office1* measuring $12m^2$ and *office2* $18m^2$, it would be more appropriate to talk about *office1* as "the smaller office," not as "the small office." The reason being that it is still bigger than the average office. These ideas are based on insights in categorization and *prototypicality* originating with Brown [7] and Rosch [53]: some instances in a category are more prototypical of that category than others.

Zender et al (§2.1.1, §2.1.2) focus on the second and third aspect, namely the problem of including the right amount of information that allows the hearer to identify the intended referent. According to the seminal work on *generating referring expressions* (GRE) by Dale and Reiter [14], one needs to distinguish whether the intended referent is already in the hearer's *focus of attention* or not. This focus of attention can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (modeled as beliefs in the belief model associated with the dialogue context). If the intended referent is already part of the current context, the GRE task merely consists of singling out the referent among the other members of the context, which act as distractors. In this case the generated referring expression (RE) contains *discriminatory* information, e.g. "the red ball" if several kinds of objects with different colors are in the current context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

While most existing GRE approaches assume that the intended referent is part of a given scene model, the *context set*, very little research has investigated the nature of references to entities that are not part of the current context. The domain of such systems is usually a small visual scene, e.g. a number of objects, such as cups and tables, located in the same room, other closed-context scenarios, including a human-robot collaborative table-top scenario [14, 31, 35, 33]. What these scenarios have in common is that they focus on a limited part of space, which is immediately and fully observable: *small-scale space*.

In contrast, mobile robots typically act in more complex environments. They operate in *large-scale space*, i.e. space "larger than what can be perceived at once" [39]. At the same time they do need the ability to understand and produce verbal references to things that are beyond the current visual and spatial context. When talking about remote places and things outside the current focus of attention, the task of *extending the context* becomes key.

Paraboni et al. [46] are among the few to address this problem. They present an algorithm for *context determination* in hierarchically ordered domains, e.g. a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g. figures and tables in book chapters), and consequently they do not touch upon the challenges that arise in a physically and perceptually situated dialogue setting. Nonetheless the approach presents a number of contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through *Ancestral Search*. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that their GRE algorithm treats spatial relationships as one-place attributes. For example a spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") is given as a property of the room entity (BUILDING NAME = COCKROFT), rather than a two-place relation (`in(library,Cockroft)`). Thereby they avoid recursive calls to the GRE algorithm, which would be necessary if the intended referent is related to another entity that needs to be properly referred to. Zender et al argue that this imposes an unnecessary restriction onto the design of the knowledge base. Moreover, it makes it hard to use their context determination algorithm as a sub-routine of any of the many existing GRE algorithms. They show how these shortcomings can be overcome, in an approach that integrates context- and content-determination as separate routines. The approach is furthermore *bi-directional*, meaning it can be used for both producing and comprehending referring expressions.

Zender & Kruijff (§2.1.4) present preliminary research on a method that enables a robot to introspect what it knows and doesn't know about an instance, relative to a given category. The method is based on the idea of querying the robot's ontological knowledge to retrieve the properties that an entity would need to fulfill to be an instance of that given category. The robot can then compare these properties to those that it already knows for the instance. Working under an open world assumption, the robot can then consider any remaining properties as gaps, indicating ignorance. This basic idea is similar to slot-filling strategies in *information states-based dialogue management* [63]. An information state is a set of records of what we would like to know, and what we already know. Any open records identify "gaps" that we need to fill next – for example, if our state reflects booking a train ticket, records may indicate departure, arrival, destination, etc. A dialogue system for booking a ticket then will ask the user for all these bits of information, to ensure it can get the user the right ticket. Here we face something similar: obtain all the information for a set of properties so that we can establish the entity as an instance of a given category. Having said that, Zender & Kruijff indicate how the method has the potential to go beyond a slot-filling strategy, in several ways. They argue how the method can extended to deal with uncertainty in categorization, and use weighted abduction of the kind proposed by Kruijff & Janiček (§2.2.3) to provide a "lowest-cost" way of establishing the right category for the entity. This again follows up on the general CogX perspective, integrating different sources of information to help overcome uncertainty in understanding (perceptual data, ontological knowledge) to drive inferences towards establishing an interpretation (weighted abduction). Zender & Kruijff extend a recent method for verbalizing ontological structure [55] to properly reflect what the robot knows about the category, the instance, and the gaps it has identified.

1.3.2 Clarification

Kruijff & Brenner (§2.2.1) propose the notion of *question nucleus*. This notion captures the information pertaining to a question. A description logic-like formalism is used to represent such information, as a conceptual structure in which propositions have ontological sorts and unique indices, and can be related through named relations. A question can then be represented as a structure in which we are querying one or more aspects of such a representation [23, 36]. The formalism allows everything to be queried: relations, propositions, sorts. The nucleus altogether comprises the situation (the "facts") and the beliefs that have led up to the question, the question itself, and the goal content which would resolve the question. The question nucleus thus integrates Ginzburg's notions of *aboutness* and *(potential) resolvedness*, and includes an explicit notion of what information is shared, and what is privately held information (cf. [42, 26]). Intuitively, it thus represents what the robot is asking about (aboutness), what it would like to know (resolvedness), and how it can appeal to shared beliefs or needs to make clear private beliefs when raising the question. The contributions the approach aims for are, briefly, as follows. Purver and Ginzburg develop an account for generating questions in a dialogue context [51, 50]. Their focus was, however, on clarification for the purpose of dialogue grounding. A similar observation can be made for recent work in HRI [41]. Kruijff & Brenner are more interested in formulating questions regarding issues in building up situation awareness, including the acquisition of new ways of understanding situations (cf. also [36]). In issue-based (or information state-based) dialogue systems [40], the problem of how to phrase a question is greatly simplified because the task domain is fixed. There is little need for paying attention to transparency or scaffolding, as it can be assumed the user understands the task domain. This is however an assumption that cannot be made for our setting.

Kruijff & Janiček (§2.2.3) provide a model for capturing the *continual* nature of *collaborative activity*. They base their approach on an algorithm in which a form of *weighted abduction* plays a core role. Weighted abduction is "inference to the best explanation" – meaning, in this context, the best explanation for why someone is saying something, and formulating that explanation in terms of an intention, an update to a belief model, and possible updates to an attentional state. Using weighted abduction for interpretation of natural language was introduced by Hobbs et al in [30]. Kruijff &

Janiček use an extended form, proposed by Stone & Thomason [60, 61, 62]. Stone & Thomason's approach integrates *attentional state*, intention, and beliefs. Their attentional state captures those entities that are currently "in focus" or highly salient in the context. (Kruijff & Janiček turn this into beliefs about such entities.) The approach is related to other collaborative models of dialogue [27, 42, 26], and provides a single model for both comprehension and production. Stone & Thomason's notion of "context" provides for a more flexible way of resolving contextual references than classical discourse theories, though. Beliefs, intentions, and attentional state can all co-determine the conditions on resolving a reference – rather than that resolution is solely determined by structural aspects of discourse (like in e.g. SDRT [2]). This provides a suitable bridge to the continuum between action and interaction, which Kruijff & Brenner have argued for, cf. Brenner et al §2.2.2. Kruijff & Janiček propose to extend Stone & Thomason's approach with a more explicit notion of situated multi-agent belief models, and they introduce assertions into proofs. An assertion is a statement whose "future necessary truth" needs to be assumed for a proof to conclude. This notion of assertion is taken from continual planning [6] where it is used to state the necessity of a future observation. Depending on the verification of the observation, an assertion triggers explicit expanding or revision of a plan. Within an abductive proof, an assertion turns the corresponding action plan into a continual plan, to achieve the inferred update to the agent's belief model and attentional state. Assertions thus make Stone & Thomason's intuitive idea of "checkpoints" more precise. Kruijff & Janiček explore the use of assertions in abductive proofs in the context of producing and comprehending clarification dialogues.

Kruijff-Korbayová et al (§2.2.4) explore intonation in situated dialogue, with a particular focus on intonation in questions like clarification requests. Intonation of clarification requests has so far received relatively little attention in the literature. Previous work on controling accent placement and type in dialogue system output based on information structure assigment w.r.t. the context all concentrated on the assignment of intonation in statements [49, 38, 4]. The seminal work of [51] which laid out a classification of the forms and functions of clarification requests based on extensive corpus analysis does not take intonation into account. Pioneering in this respect is the study of CRs in German task-oriented human-human dialogues in [52], who found that the use of intonation seemed to disambiguate clarification types, with rising boundary tones used more often to clarify acoustic problems than to clarify reference resolution. A series of production and perception experiments with one-word grounding utterances in Swedish has also shown differences in prosodic features depending on meaning (acknowledgment vs. clarification of understanding or perception), and that subjects differentiate between the meanings accordingly, and respond differently [17, 58]. The work by Kruijff-Korbayová et al extends the use of information structure to control the intonation of dialogue system output beyond answers to information-seeking questions: they include acknowledgments as well as clarification requests, and ultimately other types of questions. They include both fragmentary grounding feedback and full utterances, and address varying placement of pitch accents depending on context and communicative intention.

1.3.3 Robust processing of spoken situated dialogue

Lison & Kruijff's work on robust processing (§2.3.1, §2.3.2) aims to address two central issues in spoken dialogue processing: (1) disfluencies in verbal interaction and (2) speech recognition errors.

We know from everyday experience that spoken language behaves quite differently from written language. We do not speak the way we write. The difference of communicative medium plays a major role in this discrepancy. A speech stream offers for instance no possibility for "backtracking" – once something has been uttered, it cannot be erased anymore. And, contrary to written language, the production of spoken language is strongly *time-pressured.* The pauses which are made during the production of an utterance do leave a trace in the speech stream. As a consequence, spoken dialogue is replete with *disfluencies* such as filled pauses, speech repairs, corrections or repetitions [56]. A speech stream is also more difficult to segment and delimitate than a written sentence with punctuation and clear empty spaces between words. In fact, the very concepts of "words" and "sentences", which are often taken as core linguistic objects, are much more difficult to define with regard to spoken language. When we analyse spoken language, we observe a continuous speech stream, not a sequence of discrete objects. Hence the presence of many *discourse markers* in spoken dialogue, which play an important role in determining discourse structure. A final characteristic of spoken dialogue which is worth pointing out is that few spoken utterances take the form of complete sentences. The most prototypical example is the "short answer" in response to queries, but many other types of fragments or *non-sentential utterances* can be found in real dialogues [19]. This is mainly due to the *interactive* nature of dialogue – dialogue participants heavily rely on what has been said previously, and seek to avoid redundancies. As a result of all these factors, spoken language contains much more disfluent, partial, elided or ungrammatical utterances than written language. The question of how to *accommodate* these types of ill-formed input is a major challenge for spoken dialogue systems.

A second, related problem is *automatic speech recognition* (ASR). Speech recognition is the first step in comprehending spoken dialogue, and a very important one. For robots operating in real-world, noisy environments, and dealing with utterances pertaining to complex, open-ended domains, this step is also highly error-prone. In spite of continuous technological advances, the performance of ASR indeed remains for most tasks at least an order of magnitude worse than that of human listeners [44]. And contrary to human performance, ASR accuracy is usually unable to *degrade gracefully* when faced with new conditions in the environment (ambient noise, bad microphone, non-native or regional accent, variations in voice intensity, etc.) [12]. This less-than-perfect performance of ASR technology seriously hinders the robustness of dialogue comprehension systems, and new techniques are needed to alleviate this problem[1].

The papers included in this deliverable present an integrated approach to dealing with these problems. The approach has three defining characteristics:

1. It is a hybrid approach, combining symbolic and statistical methods to process spoken dialogue. The implemented mechanisms combine fine-grained linguistic resources (a CCG lexicon) with statistical information (the ASR language model and the discriminative model). The resulting system therefore draws from the best of both worlds and is able to deliver both *deep* and *robust* language processing.

2. It is also an integrated approach to spoken dialogue comprehension. It goes all the way from the signal processing of the speech input up to the logical forms and the pragmatic interpretation. The various components involved in dialogue processing interact with each other in complex ways to complement, coordinate and constrain their internal representations.

3. Finally, it is also a context-sensitive approach. Contextual information is used at each processing step, either as an *anticipatory* mechanism (to guide expectations about what is likely to be uttered next), or as a *discriminative* mechanism (to prune interpretations which are contextually unlikely). These mechanisms are implemented by the dynamic adaptation of the ASR language model and the use of contextual features in the discriminative model for robust parsing.

This approach compares to the state of the art in robust processing of spoken dialogue, as follows. Commercial spoken dialogue systems traditionally rely on shallow parsing techniques such as "concept spotting". In this approach, a small hand-crafted, task-specific grammar is used to extract specific constituents, such as locative phrases or temporal expressions, and turn these into basic semantic concepts [65, 32, 3, 16, 1]. These techniques are usually very efficient, but also present several important shortcomings,

[1] The speech recogniser included into our robotic platform – Nuance Recognizer v8.5 with statistical language models – yields for instance a word error rate (WER) of about 20 % when evaluated on real spoken utterances. Thus, more than *one word out of five* in each utterance is actually misrecognised by the system.

as they are often highly domain-specific, fragile, and require a lot of development and optimisation effort to implement. In more recent years, several new techniques emerged, mainly based on statistical approaches. In the CHORUS system [47], the utterances are modeled as Hidden Markov Models [HMMs], in which hidden states correspond to semantic concepts and the state outputs correspond to the individual words. HMMs are however a flat-concept model – the semantic representation is just a linear sequence of concepts with no internal structure. To overcome this problem, various stochastic parsing techniques have been proposed, based either on Probabilistic Context Free Grammars [43, 20], lexicalised models [13, 10], data-oriented parsing [5, 57], or constrained hierarchical models [29]. A few recent systems, such as the SOUP parser, also attempt to combine shallow parsing with statistical techniques, based on a hand-crafted grammar associated with probabilistic weights [22]. More rarely, we can also find in the literature some descriptions of spoken dialogue systems performing a real grammatical analysis, usually along with a “robustness” mechanism to deal with speech recognition errors, extra-grammaticality [64, 9] or ill-formed inputs [66].

Compared to the state of the art, our approach is unique in the sense that it is, to the best of our knowledge, the only one which attempts to combine deep grammatical analysis together with statistical discriminative models exploiting both linguistic and contextual information. This has arguably several advantages. Using a deep processing approach, we are able to extract full, detailed semantic representations, which can then be used to draw inferences and perform sophisticated dialogue planning. This is not possible with shallow or statistical methods. At the same time, due to the grammar relaxation mechanism and the discriminative model, we do not suffer from the inherent fragility of purely symbolic methods. Our parsing method is particularly robust, both to speech recognition errors and to ill-formed utterances. Finally, contrary to “concept spotting” techniques, our approach is much less domain-specific: the parser relies on a general-purpose lexicalised grammar which can be easily reused in other systems.

Our approach is also original in its tight integration of multiple knowledge sources – and particularly contextual knowledge sources – all through the utterance comprehension process. Many dialogue systems are designed in a classical modular fashion, where the output of a component serves as direct input for the next component, with few or no interactions other than this pipelined exchange of data[2]. Our strategy, however, is to put the tight, multi-level integration of linguistic and contextual information at the very center of processing.

As a final note, we would like to stress that our dialogue comprehension system also departs from previous work in the way we define “context”.

[2] Some interesting exceptions to this design include integrated approaches such as [45, 21].

Many recent techniques have been developed to take context into account in language processing (see e.g. [28]). But the vast majority of these approaches take a rather narrow view of context, usually restricting it to the mere dialogue/discourse context. Our dialogue comprehension system is one of the only ones (with the possible exceptions of [54, 8, 25]) to define context in a multimodal fashion, with a special focus on situated context.

2 Annexes

2.1 Verbalization

2.1.1 Zender et al. "A Situated Context Model for Resolution and Generation of Referring Expressions" (ENLG'09)

Bibliography H. Zender, G.J.M. Kruijff, and I. Kruijff-Korbayová. "A Situated Context Model for Resolution and Generation of Referring Expressions." In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009). pp. 126–129. Athens, Greece. March 2009.

Abstract The background for this paper is the aim to build robotic assistants that can naturally interact with humans. One prerequisite for this is that the robot can correctly identify objects or places a user refers to, and produce comprehensible references itself. As robots typically act in environments that are larger than what is immediately perceivable, the problem arises how to identify the appropriate context, against which to resolve or produce a referring expression (RE). Existing algorithms for generating REs generally by-pass this problem by assuming a given context. In this paper, we explicitly address this problem, proposing a method for context determination in large-scale space. We show how it can be applied both for resolving and producing REs.

Relation to WP The paper makes it possible for the robot to discuss objects and places beyond the currently perceivable situation. That makes it unnecessary for a robot and a human to be in the very place where there is something a robot needs to be explained.

2.1.2 Zender et al. "Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants." (IJCAI'09)

Bibliography H. Zender, G.J.M. Kruijff, and I. Kruijff-Korbayová. "Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants." In: Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09). Pasadena, CA, USA. July 2009.

Abstract In this paper we present an approach to the task of generating and resolving referring expressions (REs) for conversational mobile robots. It is based on a spatial knowledge base encompassing both robot- and human-centric representations. Existing algorithms for the generation of referring expressions (GRE) try to find a description that uniquely identifies the referent with respect to other entities that are in the current context. Mobile robots, however, act in large-scale space, that is environments that are larger than what can be perceived at a glance, e.g. an office building with different floors, each containing several rooms and objects. One challenge when referring to elsewhere is thus to include enough information so that the interlocutors can extend their context appropriately. We address this challenge with a method for context construction that can be used for both generating and resolving REs two previously disjoint aspects. Our approach is embedded in a bi-directional framework for natural language processing for robots.

Relation to WP The paper further explores how a robot can discuss objects and places outside the current situation (cf. also §2.1.1). The paper shows how determining the appropriate context for a reference can be integrated in a bi-directional approach, to enable the robot to both produce and comprehend contextually appropriate references.

2.1.3 Zender and Pronobis. "Verbalizing vague scalar predicates for autonomously acquired ontological knowledge" (report)

Bibliography H. Zender and A. Pronobis. "Verbalizing vague scalar predicates for autonomously acquired ontological knowledge" (report)

Abstract The paper reports on ongoing research in generating and understanding verbal references to entities in the robot's environment. The paper focuses on features of spatial entities that are commonly expressed as vague scalar predicates in natural language, such as, e.g., size. The paper proposes an approach for characterizing such features in terms of properties and distributions over their values. This leads to a basic notion of prototypicality of property-values. Using this notion, the paper shows how different types of contextual standards can be defined, which determine the contextually appropriate use of a vague scalar predicate in linguistically describing a feature of a spatial entity. The approach goes beyond existing work in that it allows for a variety of contextual standards (in class, across classes, across instances) in describing features as vague scalar predicates, and by ultimately basing these standards in models of the robot's perceptual experience.

Relation to WP Typically a robot is provided with an initial ontology, outlining the concepts and their relations considered relevant for understanding the environment in which the robot is to act. Over time, the robot can extend this ontology, for example with instances and properties that hold for these. The report develops a new method for verbalizing knowledge about autonomously acquired scalar properties for instances and their classes.

2.1.4 Zender and Kruijff. "Verbalizing classes and instances in ontological knowledge" (report)

Bibliography H. Zender and G.J.M. Kruijff. "Verbalizing classes and instances in ontological knowledge" (report)

Abstract The paper reports preliminary research on verbalizing a robot's knowledge about an instance I of a particular category C. This covers both what a robot knows, and what it does not (yet) know about the instance. The paper considers a "gap" to be that information the robot misses to establish a given property P for I, knowing that that property typically applies to instances of C. The paper proposes a method for determining which properties are classifiable as gaps for an instance relative to a category. This method operates on the T- and A-box of an ontology. It provides a general method for determining gaps, and is not specific to situated dialogue. The paper shows how the resulting characterization of available and missing knowledge about I relative to C can then be verbalized, following up an approach recently presented in [55]. The paper illustrates the method on an example involving spatial entities, and discusses further research on extending the method.

Relation to WP The report provides a first attempt at verbalizing ontological knowledge about classes and instances, with a particular focus on verbalizing what a robot does not yet know about a particular instance (i.e. a "gap").

2.2 Clarification

2.2.1 Kruijff and Brenner. "Phrasing Questions" (AAAI SS'09)

Bibliography G.J.M. Kruijff and M. Brenner. "Phrasing Questions." In: Proceedings of the AAAI 2009 Spring Symposium on Agents that Learn from Human Teachers. Stanford, CA. March 2009.

Abstract In a constructive learning setting, a robot builds up beliefs about the world by interacting – interacting with the world, and with other agents. Asking questions is key in such a setting. It provides a mechanism for interactively exploring possibilities, to extend and explain the robot's beliefs. The paper focuses on how to linguistically phrase questions in dialogue. How well the point of a question gets across depends on how it is put. It needs to be effective in making transparent the agent's intentions and beliefs behind raising the question, and in helping to scaffold the dialogue such that the desired answers can be obtained. The paper proposes an algorithm for deciding what to include in formulating a question. Its formulation is based on the idea of considering transparency and scaffolding as referential aspects of a question.

Relation to WP The paper considers what beliefs to use as context for a question (considered as a subdialogue). The paper defines the notion of a question nucleus. This structure identifies beliefs that provide a background for the question, the expected answers to the question, and a plan for formulating the question. The identified beliefs provide the basis for determining how to achieve transparency in phrasing the question, by relating aspects of the question nucleus to private and shared beliefs.

2.2.2 Brenner et al. "Continual Collaborative Planning for Situated Interaction" (report)

Bibliography M. Brenner, G.J.M. Kruijff, I. Kruijff-Korbayová, and N.A. Hawes. "Continual Collaborative Planning for Situated Interaction."

Abstract When several agents are situated in a common environment they usually interact both verbally and physically. Human-Robot Interaction (HRI) is a prototypical case of such situated interaction. It requires agents to closely integrate dialogue with behavior planning, physical action execution, and perception. The paper describes a framework called Continual Collaborative Planning (CCP) and its application to HRI. CCP enables agents to autonomously plan and realise situated interaction that intelligently interleaves planning, acting, and communicating. The paper analyses the behavior and efficiency of CCP agents in simulation, and on two robot implementations.

Relation to WP The paper argues for the continual nature of dialogue processing, reacting to the dynamics of the collaborative activity encompassing the actions of the different agents, and their interaction.

2.2.3 Kruijff and Janiček. "Abduction for clarification in situated dialogue" (report)

Bibliography G.J.M. Kruijff and M. Janiček. "Abductive inference for clarification in situated dialogue" (report)

Abstract A robot can use situated dialogue with a human, in an effort to learn more about the world it finds itself in. When asking the human for more information, it needs to be clear to the human, what the robot is talking about. The robot needs to make transparent what it would like to know more about, what it does know (or doesn't), and what it is after. Otherwise, the human is less likely to provide a useful answer to the robot. They need to establish a common ground in. The paper presents ongoing research on developing an approach for comprehending and producing (sub-)dialogues for clarifying or requesting information about the world in which establishing common ground in beliefs, intentions, and attention plays an explicit role. The approach is based on Stone & Thomason's abductive framework [60, 61, 62]. This framework integrates intention, attentional state, and dynamic interpretation to abductively derive an explanation on what assumptions and intentions communicated content can be interpreted as updating a belief context. The approach extends the framework of Stone & Thomason with assertions, to provide an explicit notion of checkpoint, and a more explicit form of multi-agent beliefs [6]. The approach uses these notions to formulate clarification as continual process of comprehension and production set in dialogue as a collaborative activity.

Relation to WP The report details a continual approach for managing clarification dialogues, based on an extended form of weighted abductive inference. The inference process covers both comprehension and production, in an interleaved fashion. The approach integrates intention, attentional state, and multi-agent belief models in a continual way of dealing with dialogue as a collaborative activity.

2.2.4 Kruijff-Korbayová et al. "Contextually appropriate intonation of clarification in situated dialogue" (report)

Bibliography I. Kruijff-Korbayová, R. Meena, and G.J.M. Kruijff. "Contextually appropriate intonation of clarification in situated dialogue." Report.

Abstract When in doubt, ask. This paradigm very much applies to autonomous robots which self-understand and self-extend in the environment they find themselves. For this, it is essentially for these systems to learn continuously, driven mainly by their own curiosity about the surroundings. Spoken dialogue is a means through which a robot can clarify or extend the acquired knowledge about the situated environment. This ability to self-initiate a dialogue to actively seek information or clarifications besides adding autonomy to a robot's behavior also allows the robot to connect its belief system to that of its listener. This access to respective belief systems in a dialogue helps the participating agents in dialogue *grounding.* However, for conversational robots raising clarification requests to seeking information is not only limited to contextually appropriate lexical selection and utterance content planning, but extends further to the generation of contextually appropriate intonation. In the absence of contextually appropriate intonation, dialogue participants might be lead to maintain incongruous belief state in wake of situational ambiguities that may arise in situated dialogue. Use of contextually appropriate intonation in clarification statements will enable the robot to rightly express its intentions to the human interlocutor. In this work we develop an approach for determining contextually appropriate intonation in clarification statements, for resolving situated ambiguities. Following the approaches [24, 50, 51] to clarification in human dialogue, we develop clarification strategies in human-robot dialogue for continuous and cross-modal learning. Working in the lines of Steedman's theory of *information structure* [59, 48] and [18], we propose and develop the notion of information packaging in our clarification statements. We evaluate our approach to generation of contextually appropriate intonations using psycholinguistically plausible experimental setup.

Relation to WP When a robot raises a question, or more in general says something in a given context, it is important for it to be clear how the utterance relates to the preceding context – and what it focuses on. Intonation is one such means to indicate this relation to context.

2.3 Robust processing of spoken situated dialogue

Increased robustness, ultimately reflected as an improvement in understanding what the user has said, contributes to efficient and effective dialogue: the better the understanding, the less need for corrective measures (e.g. clarification).

2.3.1 Lison and Kruijff. "An integrated approach to robust processing of situated spoken dialogue." (SRSL'09)

Bibliography P. Lison and G.J.M. Kruijff. "An integrated approach to robust processing of situated spoken dialogue." In: Proceedings of the Second International Workshop on the Semantic Representation of Spoken Language (SRSL'09). Athens, Greece. April 2009

Abstract Spoken dialogue is notoriously hard to process with standard NLP technologies. Natural spoken dialogue is replete with disfluent, partial, elided or ungrammatical utterances, all of which are difficult to accommodate in a dialogue system. Furthermore, speech recognition is known to be a highly error-prone task, especially for complex, open-ended domains. The combination of these two problems - ill-formed and/or misrecognised speech inputs - raises a major challenge to the development of robust dialogue systems. We present an integrated approach for addressing these two issues, based on an incremental parser for Combinatory Categorial Grammar. The parser takes word lattices as input and is able to handle ill-formed and misrecognised utterances by selectively relaxing its set of grammatical rules. The choice of the most relevant interpretation is then realised via a discriminative model augmented with contextual information. The approach is fully implemented in a dialogue system for autonomous robots. Evaluation results on a Wizard of Oz test suite demonstrate very significant improvements in accuracy and robustness compared to the baseline.

Relation to WP The paper describes an approach in which context information (salient entities, properties, and actions) is used to anticipate likely word sequences (biasing the lexical activations of words in a language model), and to discriminate (complete) parses. This yields improvements in robustness, resulting in a lower word error rate (WER) and an improvement in partial- and exact-matches of semantic representations against a WoZ corpus.

2.3.2 Lison and Kruijff, "Efficient parsing of spoken inputs for human-robot interaction" (RO-MAN'09)

Bibliography P. Lison and G.J.M. Kruijff. "Efficient parsing of spoken inputs for human-robot interaction." In: Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09). Toyama, Japan. September 2009.

Abstract The use of deep parsers in spoken dialogue systems is usually subject to strong performance requirements. This is particularly the case in human-robot interaction, where the computing resources are limited and must be shared by many components in parallel. A real-time dialogue system must be capable of responding quickly to any given utterance, even in the presence of noisy, ambiguous or distorted input. The parser must therefore ensure that the number of analyses remains bounded at every processing step. The paper presents a practical approach to address this issue in the context of deep parsers designed for spoken dialogue. The approach is based on a word lattice parser combined with a statistical model for parse selection. Each word lattice is parsed incrementally, word by word, and a discriminative model is applied at each incremental step to prune the set of resulting partial analyses. The model incorporates a wide range of linguistic and contextual features and can be trained with a simple perceptron. The approach is fully implemented as part of a spoken dialogue system for human-robot interaction. Evaluation results on a Wizard-of-Oz test suite demonstrate significant improvements in parsing time.

Relation to WP Whereas the (SRSL'09) paper only considers the uses of discriminative models at the end of the parsing process, the current paper employs discriminative models after each incremental step during parsing. A discriminative models ranks all partial analyses, after which the top-30 ranked analyses are selected for further processing. The paper shows a 50% improvement in parsing time, without any significant loss in performance (partial/exact match). Improvements in processing time make it possible for the system to have a faster response-time.

References

[1] J.F. Allen, B.W. Miller, E.K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *ACL'96: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, USA, 1996. Association for Computational Linguistics.

[2] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, Cambridge, UK, 2003.

[3] Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. The philips automatic train timetable information system. *Speech Communications*, 17(3-4):249–262, 1995.

[4] Rachel Baker, Robert A. J. Clark, and Michael White. Synthetizing contextually appropriate intonation in limited domains. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004.

[5] Rens Bod. Context-sensitive spoken dialogue processing with the dop model. *Natural Language Engineering*, 5(4):309–323, 1999.

[6] M. Brenner and B. Nebel. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*, 2008.

[7] Roger Brown. How shall a thing be called? *Psychological Review*, 65(1):14–21, 1958.

[8] J. Y. Chai and Sh. Qu. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing 2005*, pages 217–224, Vancouver, Canada, October 2005. Association for Computational Linguistics.

[9] J.-P. Chanod. Robust parsing and beyond. In Gertjan van Noord and J. Juncqua, editors, *Robustness in Language Technology*. Kluwer, 2000.

[10] Eugene Charniak. Immediate-head parsing for language models. In *ACL '01: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 124–131, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[11] H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.

[12] Ronald A. Cole and Victor Zue. Spoken language input. In Ronald A. Cole, Joseph Mariana, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, 1997.

[13] M. Collins. Three generative, lexicalised models for statistical parsing. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[14] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[15] D. DeVault and M. Stone. Interpreting vague utterances in context. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 2004.

[16] John Dowding, Robert Moore, Francois Andry, and Douglas Moran. Interleaving syntax and semantics in an efficient bottom-up parser. In *ACL-94: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 110–116. Association for Computational Linguistics, 1994.

[17] Jens Edlund, Davod House, and Gabriel Skantze. The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech. Lisbon, Portugal*, pages 2389–2392, 2005.

[18] E. Engdahl. Infromation packaging in questions. *Empirical Issues in Syntax and Semantics*, 6(1):93–111, 2006.

[19] R. Fernández and J. Ginzburg. A corpus study of non-sentential utterances in dialogue. *Traitement Automatique des Langues*, 43(2):12–43, 2002.

[20] Shai Fine. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[21] Malte Gabsdil and Johan Bos. Combining acoustic confidence scores with deep semantic analysis for clarification dialogues. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5*, pages 137–150, 2003.

[22] Marsal Gavaldà. Soup: a parser for real-world spontaneous speech. In *New developments in parsing technology*, pages 339–350. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[23] J. Ginzburg. The semantics of interrogatives. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory*. Blackwell, 1995.

[24] J. Ginzburg. Interrogatives: Questions, facts and dialogue. In *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell, 1996.

[25] P. Gorniak and D. Roy. Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2):197–231, 2007.

[26] B.J. Grosz and S. Kraus. The evolution of shared plans. In A. Rao and M. Wooldridge, editors, *Foundations and Theories of Rational Agency*, pages 227–262. Springer, 1999.

[27] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[28] A. Gruenstein, C. Wang, and S. Seneff. Context-sensitive statistical language modeling. In *Proceedings of INTERSPEECH 2005*, pages 17–20, 2005.

[29] Yulan He and S. Young. Semantic processing using the hidden vector state model. computer speech and language. *Computer Speech and Language*, 19:85–106, 2005.

[30] J.R. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

[31] Helmut Horacek. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL*, Madrid, Spain, 1997.

[32] Eric Jackson, Douglas Appelt, John Bear, Robert Moore, and Ann Podlozny. A template matcher for robust nl interpretation. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 190–194, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[33] J. Kelleher and G.J.M. Kruijff. Incremental generation of spatial referring expressions in situated dialogue. In *Proc. Coling-ACL-2006*, Sydney, Australia, 2006.

[34] C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, February 2007.

[35] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing.* CSLI Publications, Stanford, CA, USA, 2002.

[36] G.J.M. Kruijff, M. Brenner, and N.A. Hawes. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, Munich, Germany, 2008.

[37] G.J.M. Kruijff, H. Zender, P. Jensfelt, and H.I. Christensen. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*, 2006.

[38] Ivana Kruijff-Korbayová, Stina Ericsson, Kepa Joseba Rodríguez, and Elena Karagjosova. Producing contextually appropriate intonation is an information-state based dialogue system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234. ACL, 2003.

[39] B. Kuipers. *Representing Knowledge of Large-scale Space.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1977.

[40] Staffan Larsson. *Issue-Based Dialogue Management.* Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden, 2002.

[41] S. Li, B. Wrede, and G. Sagerer. A computational model of multi-modal grounding. In *Proceedings of the ACL SIGdial workshop on discourse and dialog*, pages 153–160, 2006.

[42] K. Lochbaum, B.J. Grosz, and C.L. Sidner. Discourse structure and intention recognition. In R. Dale, H. Moisl, , and H. Somers, editors, *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text.* Marcel Dekker, New York, 1999.

[43] Scott Miller, Richard Schwartz, Robert Bobrow, and Robert Ingria. Statistical language processing using hidden understanding models. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 278–282, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[44] R. K. Moore. Spoken language processing: piecing together the puzzle. *Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing*, 49:418–435, 2007.

[45] Robert Moore, John Dowding, J. M. Gawron, and Douglas Moran. Combining linguistic and statistical knowledge sources in natural-language processing for atis. In *ARPA Spoken Language Technology Workshop*, 1995.

[46] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June 2007.

[47] Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, Esther Levin, Chin-Hui Lee, and Jean-Luc Gauvain. Progress report on the Chronus system: ATIS benchmark results. In *HLT '91: Proceedings*

of the workshop on Speech and Natural Language, pages 67–71, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[48] Scott A. Prevost. An information structural approach to spoken language generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 294–301, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[49] Scott A. Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation.* Phd thesis, University of Pennsylvania, Institute for Research in Cognitive Science Technical Report, Pennsylvania, USA, 1996.

[50] M. Purver. *The Theory and Use of Clarification Requests in Dialogue.* PhD thesis, King's College, University of London, 2004.

[51] M. Purver, J. Ginzburg, and P. Healey. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers, 2003.

[52] Kepa J. Rodríguez and David Schlangen. Form, intonation and function of clarification requests in german task oriented spoken dialogues. In *Proceedings of Catalog '04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04), Barcelona, Spain, July* , 2004.

[53] Eleanor Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1978.

[54] Deb Roy. Situation-aware spoken language processing. In *Royal Institute of Acoustics Workshop on Innovation in Speech Processing*, Stratford-upon-Avon, England, 2001.

[55] Niels Schütte. Generating natural language descriptions of ontology concepts. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 106–109, Athens, Greece, March 2009.

[56] E. Shriberg. Disfluencies in switchboard. In *Proceedings of ICSLP '96*, volume supplement, Philadelphia, PA, 1996.

[57] Khalil Sima'an. Robust data oriented spoken language understanding. In *New developments in parsing technology*, pages 323–338. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[58] Gabriel Skanze, David House, and Jens Edlund. User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech ICSLP. Pittsburgh PA, USA*, pages 2002–2005, 2006.

[59] M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689, 2000.

[60] M. Stone and R.H. Thomason. Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[61] M. Stone and R.H. Thomason. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*, 2003.

[62] R.H. Thomason, M. Stone, and D. DeVault. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In D. Byron, C. Roberts, and S. Schwenter, editors, *Presupposition Accommodation.* (to appear).

[63] David Traum and Staffan Larsson. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[64] G. van Noord, G. Bouma, R. Koeling, and M.-J. Nederhof. Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 1999.

[65] Wayne Ward. Understanding spontaneous speech. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 137–141, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[66] L. S. Zettlemoyer and M. Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, 2007.

A Situated Context Model for Resolution and Generation of Referring Expressions

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

The background for this paper is the aim to build robotic assistants that can "naturally" interact with humans. One prerequisite for this is that the robot can correctly identify objects or places a user refers to, and produce comprehensible references itself. As robots typically act in environments that are larger than what is immediately perceivable, the problem arises how to identify the appropriate context, against which to resolve or produce a referring expression (RE). Existing algorithms for generating REs generally bypass this problem by assuming a given context. In this paper, we explicitly address this problem, proposing a method for context determination in large-scale space. We show how it can be applied both for resolving and producing REs.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help users perform daily chores. Autonomous vacuum cleaners have already found their way into people's homes, but it will still take a while before fully conversational robot "gophers" will assist people in more demanding everyday tasks. Imagine a robot that can deliver objects, and give directions to visitors on a university campus. This robot must be able to verbalize its knowledge in a way that is understandable by humans.

A conversational robot will inevitably face situations in which it needs to refer to an entity (an object, a locality, or even an event) that is located somewhere outside the current scene, as Figure 1 illustrates. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most human-robot dialogues. Consider the following set of examples:

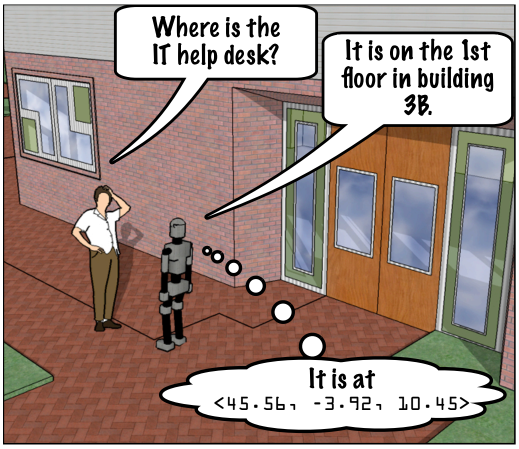

Where is the IT help desk?
It is on the 1st floor in building 3B.
It is at <45.56, -3.92, 10.45>

Figure 1: Situated dialogue with a service robot

1. "position $P = \langle 45.56, -3.92, 10.45 \rangle$"
2. "Peter's office no. 200 at the end of the corridor on the third floor of the Acme Corp. building 3 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"
3. "the area"

These REs are valid descriptions of their respective referents. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information that the hearer needs to uniquely identify the referent. The next REs *might* serve as more appropriate variants of the previous examples (*in certain contexts!*):

1. "the IT help desk"
2. "Peter's office"
3. "the large hall on the first floor"

The first example highlights a requirement on the knowledge representation to which an algorithm for generating referring expressions (GRE) has access. Although the robot needs a robot-centric representation of its surrounding space that allows it to safely perform actions and navigate its world, it should use human-centric qualitative descriptions when talking about things in the world. We do not address this issue here, but refer the interested reader to our recent work on multi-layered spatial maps for robots, bridging the gap between robot-centric and human-centric spatial representations (Zender et al., 2008).

The other examples point out another important consideration: how much information does the human need to single out the intended referent among the possible entities that the robot could be referring to? According to the seminal work on GRE by Dale and Reiter (1995), one needs to distinguish whether the intended referent is already in the hearer's *focus of attention* or not. This focus of attention can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (dialogue context). If the referent is already part of the current context, the GRE task merely consists of singling it out among the other members of the context, which act as distractors. In this case the generated RE contains *discriminatory* information, e.g. "the red ball" if several kinds of objects with different colors are in the context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

In the following we propose an approach for context determination and extension that allows a mobile robot to produce and interpret REs to entities outside the current visual context.

2 Background

Most GRE approaches are applied to very limited, visual scenes – so-called *small-scale space*. The domain of such systems is usually a small visual scene, e.g. a number of objects, such as cups and tables, located in the same room), or other closed-context scenarios (Dale and Reiter, 1995; Horacek, 1997; Krahmer and Theune, 2002). Recently, Kelleher and Kruijff (2006) have presented an incremental GRE algorithm for situated dialogue with a robot about a table-top setting, i.e. also about small-scale space. In all these cases, the context set is assumed to be identical to the visual scene that is shared between the interlocutors. The intended referent is thus already in the hearer's *focus of attention*.

In contrast, robots typically act in *large-scale space*, i.e. space "larger than what can be perceived at once" (Kuipers, 1977). They need the ability to understand and produce references to things that are beyond the current visual and spatial context. In any situated dialogue that involves entities beyond the current focus of attention, the task of *extending the context* becomes key.

Paraboni et al. (2007) present an algorithm for *context determination* in hierarchically ordered domains, e.g. a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g. figures, tables in book chapters). Consequently they do not address the challenges that arise in physically and perceptually situated dialogues. Still, the approach presents a number of good contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through Ancestral Search. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that spatial relationships are treated as one-place attributes by their GRE algorithm. For example they transform the spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") into a property of the room entity (BUILDING NAME = COCKROFT) and not a two-place relation (`in(library,Cockroft)`). Thus they avoid recursive calls to the algorithm, which would be needed if the intended referent is related to another entity that needs to be properly referred to.

However, according to Dale and Reiter (1995), these related entities do not necessarily serve as discriminatory information. At least in large-scale space, in contrast to a document structure that is conceivably transparent to a reader, they function as *attention-directing elements* that are introduced to build up *common ground* by incrementally extending the hearer's focus of attention. Moreover, representing some spatial relations as two-place predicates between two entities and some as one-place predicates is an arbitrary decision.

We present an approach for context determination (or *extension*), that imposes less restrictions on its knowledge base, and which can be used as a sub-routine in existing GRE algorithms.

3 Situated Dialogue in Large-Scale Space

Imagine the situation in Figure 1 did not take place somewhere on campus, but rather inside building 3B. Certainly the robot would not have said "the IT help desk is on the 1st floor in building 3B." To avoid confusing the human, an utterance like "the IT help desk is on the 1st floor" would have been appropriate. Likewise, if the IT help desk happened to be located on another site of the university, the robot would have had to identify its location as being "on the 1st floor in building 3B on the new campus." The hierarchical representation of space that people are known to assume (Cohn and Hazarika, 2001), reflects upon the choice of an appropriate context when producing REs.

In the above example the physical and spatial situatedness of the dialogue participants play an important role in determining which related parts of space come into consideration as potential distractors. Another important observation concerns the verbal behavior of humans when talking about remote objects and places during a complex dialogue (i.e. more than just a question and a reply). Consider the following example dialogue:

> Person A: "Where is the exit?"
>
> Person B: "You first go down this corridor. Then you turn right. After a few steps you will see the big glass doors."
>
> Person A: "And the bus station? Is it to the left?"

The dialogue illustrates how utterances become grounded in previously introduced discourse referents, both temporally and spatially. Initially, the physical surroundings of the dialogue partners form the context for anchoring references. As a dialogue unfolds, this point can conceptually move to other locations that have been explicitly introduced. Discourse markers denoting spatial or temporal cohesion (e.g. "then" or "there") can make this move to a new anchor explicit, leading to a "mental tour" through large-scale space.

We propose a general principle of *Topological Abstraction* (TA) for context extension which is rooted in what we will call the *Referential Anchor* a.[1] TA is designed for a multiple abstraction hierarchy (e.g. represented as a lattice structure rather than a simple tree). The Referential Anchor a, corresponding to the current focus of attention, forms the nucleus of the context. In the simple case, a corresponds to the hearer's physical location. As illustrated above, a can also move along the "spatial progression" of the most salient discourse entity during a dialogue. If the intended referent is outside the current context, TA extends the context by incrementally ascending the spatial abstraction hierarchy until the intended referent is an element of the resulting sub-hierarchy, as illustrated in Figure 2. Below we describe two instantiations of the TA principle, a TA algorithm for reference generation (TAA1) and TAA2 for reference resolution.

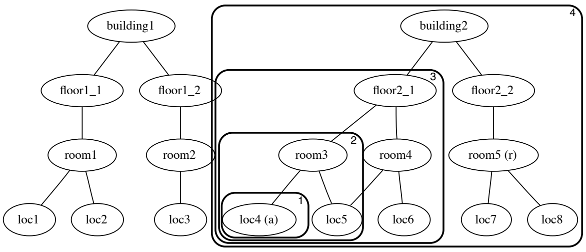


Figure 2: Incremental TA in large-scale space

Context Determination for GRE TAA1 constructs a set of entities dominated by the Referential Anchor a (and a itself). If this set contains the intended referent r, it is taken as the current utterance context set. Else TAA1 moves up one level of abstraction and adds the set of all child nodes to the context set. This loop continues until r is in the context set. At that point TAA1 stops and returns the constructed context set (cf. Algorithm 1).

TAA1 is formulated to be neutral to the kind of GRE algorithm that it is used for. It can be used with the original Incremental Algorithm (Dale and Reiter, 1995), augmented by a recursive call if a relation to another entity is selected as a discriminatory feature. It could in principle also be used with the standard approach to GRE involving relations (Dale and Haddock, 1991), but we agree with Paraboni et al. (2007) that the mutually qualified references that it can produce[2] are not easily resolvable if they pertain to circumstances where a confirmatory search is costly (such as in large-scale space). More recent approaches to avoiding infinite loops when using relations in GRE make use of a graph-based knowledge representation (Krahmer et al., 2003; Croitoru and van Deemter, 2007). TAA1 is compatible with these approaches, as well as with the salience based approach of (Krahmer and Theune, 2002).

[1] similar to Ancestral Search (Paraboni et al., 2007)

[2] An example for such a phenomenon is the expression "the ball on the table" in a context with several tables and several balls, but of which only one is on a table. Humans find such REs natural and easy to resolve in visual scenes.

Algorithm 1 TAA1 (for reference generation)

```
Require: $a$ = referential anchor; $r$ = intended referent
  Initialize context: $C = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  if $r \in C$ then
    return $C$
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ do
      for each $p \in topologicalParents(n)$ do
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      if $r \in C$ then
        return $C$
      end if
    end for
    return failure
  end if
```

Algorithm 2 TAA2 (for reference resolution)

```
Require: $a$ = ref. anchor; $desc(x)$ = description of referent
  Initialize context: $C = \{\}$
  Initialize possible referents: $R = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  $R = desc(x) \cap C$
  if $R \neq \{\}$ then
    return $R$
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ do
      for each $p \in topologicalParents(n)$ do
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      $R = desc(x) \cap C$
      if $R \neq \{\}$ then
        return $R$
      end if
    end for
    return failure
  end if
```

Resolving References to Elsewhere Analogous to the GRE task, a conversational robot must be able to understand verbal descriptions by its users. In order to avoid overgenerating possible referents, we propose TAA2 (cf. Algorithm 2) which tries to select an appropriate referent from a relevant subset of the full knowledge base. It is initialized with a given semantic representation of the referential expression, $desc(x)$, in a format compatible with the knowledge base. Then, an appropriate entity satisfying this description is searched for in the knowledge base. Similarly to TAA1, the description is first matched against the current *context set* C consisting of a and its child nodes. If this set does not contain any instances that match $desc(x)$, TAA2 increases the context set along the spatial abstraction axis until at least one possible referent can be identified within the context.

4 Conclusions and Future Work

We have presented two algorithms for context determination that can be used both for resolving and generating REs in large-scale space.

We are currently planning a user study to evaluate the performance of the TA algorithms. Another important item for future work is the exact nature of the spatial progression, modeled by "moving" the referential anchor, in a situated dialogue.

Acknowledgments

This work was supported by the EU FP7 ICT Project "CogX" (FP7-ICT-215181).

References

A. G. Cohn and S. M. Hazarika. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29.

M. Croitoru and K. van Deemter. 2007. A conceptual graph approach to the generation of referring expressions. In *Proc. IJCAI-2007*, Hyderabad, India.

R. Dale and N. Haddock. 1991. Generating referring expressions involving relations. In *Proc. of the 5th Meeting of the EACL*, Berlin, Germany, April.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

H. Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL*, Madrid, Spain.

J. Kelleher and G.-J. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialogue. In *In Proc. Coling-ACL 06*, Sydney, Australia.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA.

E. Krahmer, S. van Erk, and A. Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1).

B. Kuipers. 1977. *Representing Knowledge of Large-scale Space*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

I. Paraboni, K. van Deemter, and J. Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June.

H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J. Kruijff, and W. Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June.

Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants[*]

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

In this paper we present an approach to the task of generating and resolving referring expressions (REs) for conversational mobile robots. It is based on a spatial knowledge base encompassing both robot- and human-centric representations. Existing algorithms for the generation of referring expressions (GRE) try to find a description that uniquely identifies the referent with respect to other entities that are in the current context. Mobile robots, however, act in large-scale space, that is, environments that are larger than what can be perceived at a glance, e.g., an office building with different floors, each containing several rooms and objects. One challenge when referring to elsewhere is thus to include enough information so that the interlocutors can extend their context appropriately. We address this challenge with a method for context construction that can be used for both generating and resolving REs – two previously disjoint aspects. Our approach is embedded in a bi-directional framework for natural language processing for robots.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help the users perform their daily chores. Although the autonomous vacuum cleaner "Roomba" has already found its way into people's homes and lives, there is still a long way until fully conversational robot "gophers" will be able to assist people in more demanding everyday tasks. For example, imagine a robot that can deliver objects and give directions to visitors on a university campus. Such a robot must be able to verbalize its knowledge in a way that is understandable by humans, as illustrated in Figure 1.

A conversational robot will inevitably face situations in which it needs to refer to an entity (e.g., an object, a locality, or even an event) that is located somewhere outside the current scene. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most human-robot dialogues. Consider the following set of examples:

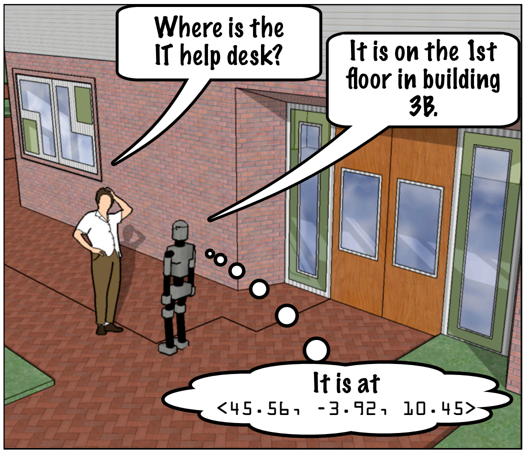

Where is the IT help desk?
It is on the 1st floor in building 3B.
It is at <45.56, -3.92, 10.45>

Figure 1: Situated dialogue with a campus service robot

1. "position $P = \langle 45.56, -3.92, 10.45\rangle$"
2. "the area"
3. "Peter's office at the end of the corridor on the third floor of the Acme Corp. building 7 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"

Clearly, these REs are valid descriptions of the respective entities in the robot's world representation. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information so that the hearer can easily uniquely identify what is meant. The following expressions *might* serve as more appropriate variants of the previous examples (*in certain situations!*):

1. "the IT help desk"
2. "the large hall on the first floor"
3. "Peter's office"

However, the question remains how a natural language processing (NLP) system can generate such expressions which are suitable in a given situation. In this paper we identify some of the challenges that an NLP system for situated dialogue about large-scale space needs to address. We present a situated model for generating and resolving REs that addresses these issues, with a special focus on how a conversational mobile robot can produce and interpret such expressions against an appropriate part of its acquired knowledge base (KB). One benefit of our approach is that most components, including the situated model and the linguistic resources, are bi-directional, i.e., they use the same representa-

[*] Supported by the EU FP7 Project "CogX" (FP7-ICT-215181).

tions for comprehension and production of utterances. This means that the proposed system is able to understand and correctly resolve all the REs that it is able to generate.

The rest of the paper is organized as follows. We first briefly discuss relevant existing approaches to comprehending and producing REs (Section 2). We then motivate our approach to context determination for situated interaction in large-scale space (Section 3), and describe its implementation in a dialogue system for an autonomous robot (Section 4). We conclude in Section 5.

2 Background

The main purpose of an RE is to enable a hearer to correctly and uniquely identify the target entity to which the speaker is referring, the so-called *intended referent*. The GRE task is thus to produce a natural language expression for a KB entity that fulfills this purpose.

As can be seen from the examples in the previous section, an RE needs to meet a number of constraints in order to be successful. First, it needs to make use of concepts that can be understood by the hearer. This becomes an important consideration when we are dealing with a robot which acquires its own models of the environment and is to talk about the contents of these. Second, it needs to contain enough information so that the hearer can distinguish the intended referent from other entities in the world, the so-called *potential distractors*. Finally, this needs to be balanced against the third constraint: Inclusion of unnecessary information should be avoided so as not to elicit false implications on the part of the hearer.

We will only briefly mention how to address the first challenge, and refer the reader to our recent work on multi-layered conceptual spatial maps for robots that bridge the gap between robot-centric representations of space and human-centric conceptualizations [Zender *et al.*, 2008].

The focus in this paper lies on the second and third aspect, namely the problem of including the right amount of information that allows the hearer to identify the intended referent. According to the seminal work on GRE by Dale and Reiter [1995], one needs to distinguish whether the intended referent is already in the hearer's *current context* or not. This context can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (dialogue context). If the intended referent is already part of the current context, the GRE task merely consists of singling out the referent among the other members of the context, which act as distractors. In this case the generated RE contains *discriminatory* information, e.g., "the red ball" if several kinds of objects with different colors are in the current context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

While most existing GRE approaches assume that the intended referent is part of a given scene model, the *context set*, very little research has investigated the nature of references to entities that are not part of the current context.

The domain of such systems is usually a small visual scene, e.g., a number of objects, such as cups and tables, located in the same room, other closed-context scenarios, including a human-robot collaborative table-top scenario [Dale and Reiter, 1995; Horacek, 1997; Krahmer and Theune, 2002; Kelleher and Kruijff, 2006]. What these scenarios have in common is that they focus on a limited part of space, which is immediately and fully observable: *small-scale space*.

In contrast, mobile robots typically act in more complex environments. They operate in *large-scale space*, i.e., space "larger than what can be perceived at once" [Kuipers, 1977]. At the same time they do need the ability to understand and produce verbal references to things that are beyond the current visual and spatial context. When talking about remote places and things outside the current focus of attention, the task of *extending the context* becomes crucial.

Paraboni et al. [2007] are among the few to address this problem. They present an algorithm for *context determination* in hierarchically ordered domains, e.g., a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g., figures and tables in book chapters), and consequently they do not touch upon the challenges that arise in a physically and perceptually situated dialogue setting. Nonetheless their approach presents a number of contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through Ancestral Search. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that their GRE algorithm treats spatial relationships as one-place attributes. E.g., a spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") is given as a property of the room entity (BUILDING NAME = COCKROFT), rather than a two-place relation (`in(library,Cockroft)`). Thereby they avoid recursive calls to the GRE algorithm, which are necessary for intended referents related to another entity that needs to be properly referred to. We claim that this imposes an unnecessary restriction onto the KB design. Moreover, it makes it hard to use their context determination algorithm as a sub-routine of any of the many existing GRE algorithms.

3 Situated Dialogue in Large-Scale Space

Imagine the situation in Figure 1 did not take place somewhere on campus, but rather inside building 3B. It would have made little or no sense for the robot to say that "the IT help desk is on the 1st floor in building 3B." To avoid confusion, an utterance like "the IT help desk is on the 1st floor" would be appropriate. Likewise, if the IT help desk happened to be located on another site of the university, the robot would have had to identify its location as being, e.g., "on the 1st floor in building 3B on the new campus". This illustrates that the hierarchical representation of space that humans adopt [Cohn and Hazarika, 2001] reflects upon the choice of an appropriate context when producing referential descriptions that involve attention-directing information.

Thus, the physical and spatial situatedness of the dialogue participants plays an important role when determining which related parts of space come into consideration as potential distractors. Another important observation concerns the verbal behavior of humans when talking about remote objects and places in a complex dialogue (i.e., more than just a question and a reply). E.g., consider the following dialogue:

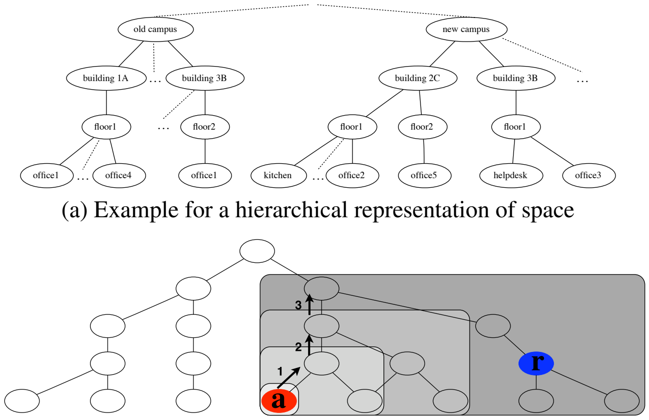
> Person A: “Where is the exit?”
> Person B: “First go down this corridor. Then turn right. After a few steps you’ll see the big glass doors.”
> Person A: “And the bus station? Is it to the left?”

As can be seen, an utterance in such a collaborative dialogue is usually grounded in previously introduced discourse referents, both temporally and spatially. Initially, the physical surroundings of the dialogue partners form the context to which references are related. Then, as the dialogue unfolds, this point can conceptually move to other locations that have been explicitly introduced. Usually, a discourse marker denoting spatial or temporal cohesion (e.g., “then” or “there”) establishes the last mentioned referent as the new anchor, creating a “mental tour” through large-scale space.

3.1 Context Determination Through Topological Abstraction

To keep track of the correct referential context in such a dialogue, we propose a general principle of *Topological Abstraction*[1] (TA) for context extension. TA is applied whenever a reference cannot be generated or resolved with respect to the current context. In such a case TA incrementally extends the context until the reference can be established. TA is designed to operate on a spatial abstraction hierarchy; i.e., a decomposition of space into parts that are related through a tree or lattice structure in which edges denote a containment relation (cf. Figure 2a). Originating in the *Referential Anchor* a, TA extends the context by incrementally ascending the spatial abstraction hierarchy until the intended referent is in the resulting sub-hierarchy (cf. Figure 2b). When no other information, e.g., from a preceding dialogue, is present, a is assumed to correspond to the spatio-visual context that is shared by the hearer and the speaker – usually their physical location and immediate surroundings. During a dialogue, however, a corresponds to the most salient discourse entity, reflecting how the *focus of attention* moves to different, even remote, places, as illustrated in the example dialogue above.

Below we describe two instantiations of the TA principle, a TA algorithm for reference generation (TAA1) and one for reference resolution (TAA2). They differ only minimally, namely in their use of an intended referent r or an RE $desc(x)$ to determine the conditions for entering and exiting the loop for topological abstraction. The way they determine a context through topological abstraction is identical.



(a) Example for a hierarchical representation of space

(b) Illustration of the TA principle: starting from the Referential Anchor (a), the smallest sub-hierarchy containing both a and the intended referent (r) is formed incrementally

Figure 2: Topological Abstraction in a spatial hierarchy

Algorithm 1 TAA1 (for reference generation)

```
Require: $a$ = referential anchor; $r$ = intended referent
  Initialize context: $C = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  if $r \in C$ then
    return $C$
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ do
      for each $p \in topologicalParents(n)$ do
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      if $r \in C$ then
        return $C$
      end if
    end for
    return failure
  end if
```

Context Determination for GRE TAA1 (cf. Algorithm 1) constructs a set of entities dominated by the Referential Anchor a (including a itself). If this set contains the intended referent r, it is taken as the current utterance context set. Else TAA1 moves up one level of abstraction and adds the set of all child nodes to the context set. This loop continues until r is in the thus constructed set. At that point TAA1 stops and returns the constructed context set.

TAA1 is formulated to be neutral to the kind of GRE algorithm that it is used for. It can be used with the original Incremental Algorithm [Dale and Reiter, 1995], augmented by a recursive call if a relation to another entity is selected as a discriminatory feature. It could in principle also be used with the standard approach to GRE involving relations [Dale and Haddock, 1991], but we agree with Paraboni et al. [2007] that the mutually qualified references that it can produce[2] are not easily resolvable if they pertain to circumstances where a confirmatory search is costly (such as in large-scale space). More recent approaches to avoiding infinite loops when using relations in GRE make use of a graph-based knowledge representation [Krahmer *et al.*, 2003; Croitoru and van Deemter, 2007]. TAA1 is compatible with these approaches, as well as with the salience based approach of Krahmer and Theune [2002].

[1] similar to Ancestral Search [Paraboni *et al.*, 2007]

[2] Stone and Webber [1998] present an approach that produces sentences like “take the rabbit from the hat” in a context with several hats and rabbits, but of which only one is in a hat. Humans find such REs natural and easy to resolve in visual scenes.

Algorithm 2 TAA2 (for reference resolution)

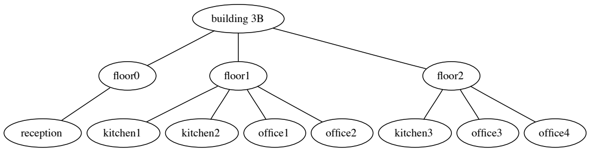
```
Require: $a$ = ref. anchor; $desc(x)$ = description of referent
  Initialize context: $C = \{\}$
  Initialize possible referents: $R = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  $R = desc(x) \cap C$
  if $R \neq \{\}$ then
    return $R$
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ do
      for each $p \in topologicalParents(n)$ do
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      $R = desc(x) \cap C$
      if $R \neq \{\}$ then
        return $R$
      end if
    end for
    return failure
  end if
```

Context Determination for Reference Resolution A conversational robot must also be able to understand verbal descriptions by its users. In order to avoid overgenerating possible referents, we propose TAA2 (cf. Algorithm 2) which tries to select an appropriate referent from a relevant subset of the full KB. It is initialized with a given semantic representation of the referential expression, $desc(x)$, in a format compatible with the KB. We will show how this is accomplished in our framework in Section 4.1. Then, an appropriate entity satisfying this description is searched for in the KB. Similarly to TAA1, the description is first matched against the current *context set* C consisting of a and its child nodes. If this set does not contain any instances that match $desc(x)$, TAA2 enlarges the context set along the spatial abstraction axis until at least one possible referent can be identified within C.

4 Implementation

Our approach for resolving and generating spatial referring expressions has been fully integrated with the dialogue functionality in a cognitive system for a mobile robot [Zender *et al.*, 2008; Kruijff *et al.*, 2009]. The robot is endowed with a *conceptual spatial map* [Zender and Kruijff, 2007], which represents knowledge about places, objects and their relations in an OWL-DL[3] ontology. We use the Jena reasoning framework[4] with its built-in OWL reasoning and rule inference facilities. Internally, Jena stores the facts of the *conceptual map* as RDF[5] triples, which can be queried through SPARQL[6] queries. Figure 3 shows a subset of such a KB.

Below, we use this example scenario to illustrate our approach to generating and resolving spatial referring expressions in the robot's dialogue system. We assume that the interaction takes place at the reception on the ground floor ("floor0"), so that for TAA1 and TAA2 a =reception.



(a) Topological abstraction hierarchy

```
(kitchen1 rdf:type Kitchen), (...)
(office1 rdf:type Office), (...)
(kitchen2 size big), (...)
(bob rdf:type Person), (bob name Bob),
(bob owns office1), (...)
(floor1 contains kitchen1), (...)
(floor2 contains office3), (...)
(floor1 ordNum 1), (floor2 ordNum 2), (...)
```

(b) RDF triples in the conceptual map (namespace URIs omitted)

Figure 3: Part of a representation of an office environment

4.1 The Comprehension Side

In situated dialogue processing, the robot needs to build up an interpretation for an utterance which is linked both to the dialogue context and to the (referenced) situated context. Here, we focus on the meaning representations.

We represent meaning as a logical form (LF) in a description logic [Blackburn, 2000]. An LF is a directed acyclic graph (DAG), with labeled edges, and nodes representing propositions. Each proposition has an ontological sort, and a unique index. We write the resulting ontologically sorted, relational structure as a conjunction of elementary predications (EPs): $@_{idx:sort}(\mathbf{prop})$ to represent a proposition **prop** with ontological sort $sort$ and index idx, $@_{idx1:sort1}\langle Rel\rangle(idx2 : srt2)$ to represent a relation Rel from index $idx1$ to index $idx2$, and $@_{idx:sort}\langle Feat\rangle(\mathbf{val})$ to represent a feature $Feat$ with value **val** at index idx. Representations are built compositionally, parsing the word lattices provided by speech recognition with a Combinatory Categorial Grammar [Lison and Kruijff, 2008]. Reversely, we use the same grammar to realize strings (cf. Section 4.2) from these meaning representations [White and Baldridge, 2003].

An example is the meaning we obtain for "the big kitchen on the first floor," (folding EPs under a single scope of @). It illustrates how each propositional meaning gets an index, similar to situation theory. "kitchen" gets one, and also modifiers like "big," "on" and "one." This enables us to single out every aspect for possible contextual reference (Figure 4a).

Next, we resolve contextual references, and determine the possible dialogue move(s) the utterance may express. Contextual reference resolution determines how we can relate the content in the utterance meaning, to the preceding dialogue context. If part of the meaning refers to previously mentioned content, we associate the identifiers of these content representations; else, we generate a new identifier. Consequently, each identifier is considered a dialogue referent.

Once we have a representation of utterance meaning in dialogue context, we build a further level of representation to facilitate connecting dialogue content with models of the robot's situation awareness. This next level of representation is essentially an a-modal abstraction over the linguistic aspects of meaning, to provide an a-modal conceptual structure

[3] http://www.w3.org/TR/owl-guide/

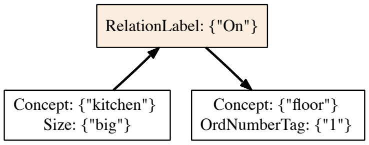
[4] http://jena.sourceforge.net

[5] http://www.w3.org/RDF

[6] http://www.w3.org/TR/rdf-sparql-query

$$\begin{array}{l}@_{l1:e-place}(\mathbf{kitchen} \wedge \\ \quad \langle Delimitation\rangle\mathbf{unique} \wedge \\ \quad \langle Num\rangle\mathbf{sg} \wedge \langle Quantification\rangle\mathbf{specific} \wedge \\ \quad \langle Modifier\rangle(b1 : q-size \wedge \mathbf{big}) \wedge \\ \quad \langle Modifier\rangle(o1 : m-location \wedge \mathbf{on} \wedge \\ \quad\quad \langle Anchor\rangle(f1 : thing \wedge \mathbf{floor} \wedge \\ \quad\quad\quad \langle Delimitation\rangle\mathbf{unique} \wedge \\ \quad\quad\quad \langle Num\rangle\mathbf{sg} \wedge \langle Quantification\rangle\mathbf{specific} \wedge \\ \quad\quad\quad \langle Modifier\rangle(n1 : number-ordinal \wedge \mathbf{1}))))\end{array}$$

(a) Logical form



(b) A-modal directed acyclic graph

```
SELECT ?x0 ?x1 WHERE {
  ?x0 rdf:type Kitchen.
  ?x0 size big.
  ?x1 rdf:type Floor.
  ?x1 ordNum 1.
  ?x0 containedIn ?x1.
```

(c) SPARQL query
In the previous example this would resolve `?x0` to `kitchen2`

Figure 4: Logical form, a-modal DAG and corresponding SPARQL query for "the big kitchen on the first floor"

[Jacobsson *et al.*, 2008]. Abstraction is a recursive translation of DAGs into DAGs, whereby the latter (conceptual) DAGs are typically flatter than the linguistic DAGs (Figure 4b).

The final step in resolving an RE is to construct a query to the robot's KB. In our implementation we construct a SPARQL query from the a-modal DAG representations (Figure 4c). This query corresponds to the logical description of the referent $desc(r)$ in TAA2. TAA2 then incrementally extends the context until at least one element of the result set of $desc(r)$ is contained within the context.

4.2 The Production Side

Production covers the entire path from handling dialogue goals to speech synthesis. The dialogue system can itself produce goals (e.g., to handle communicative phenomena like greetings), and it accepts goals from a higher level planner. Once there is a goal, an utterance content planner produces a content representation for achieving that goal, which the realizer then turns into one or more surface forms to be synthesized. Below we focus on utterance content planning.

A dialogue goal specifies a goal to be achieved, and any content that is associated with it. A typical example is to convey an answer to a user: the goal is to tell, the content is the answer. Content is given as a conceptual structure, *proto LF*, abstracting away from linguistic specifics, similar to the a-modal structures we produce for comprehension.

Content planning turns this proto LF into an LF which matches the specific linguistic structures defined in the grammar we use to realize it. "Turning into" means extending the proto LF with further semantic structure. This may be nonmonotonic in that parts of the proto LF may be rewritten, expanding into locally connected graph structures.

Planning is agenda-based, and uses a planning domain defined as a (systemic) grammar network alike [Bateman, 1997; Kruijff, 2005]. A grammar network is a collection of systems that define possible sequences of operations to be performed on a node with characteristics matching the applicability conditions for the system. A system's decision tree determines which operations are to be applied. Decisions are typically context-sensitive, based on information about the shape of the (entire) LF, or on information in context models (dialogue or otherwise). While constructing an LF, the planner cycles over its nodes, and proposes new agenda items for nodes which have not yet been visited. An agenda item consists of the node, and a system which can be applied to that node.

A system can explicitly trigger the generation of an RE for the node on which it operates. It then provides the dialogue system with a request for an RE, with a pointer to the node in the (provided) LF. The dialogue system resolves this request by submitting it to GRE modules which have been registered with the system. (Registration allows us to plug-and-play with content-specific GRE algorithms.) Assuming a GRE module produces an LF with the content for the RE, the planner gets this LF and integrates it into the overall LF.

For example, say the robot in our previous example is to answer the question "Where is Bob?". We receive a communicative goal (see below) to inform the user, specifying the goal as an assertion related to the previous dialogue context as an answer. The content is specified as an ascription e of a property to a target entity. The target entity is t which is specified as a person called "Bob" already available in the dialogue context, and thus familiar to the hearer. The property is specified as topological inclusion (TopIn) within the entity k, the reference to which is to be produced by the GRE algorithm (hence the type "rfx" and the "RefIndex" which is the address of the entity).

$$\begin{array}{l}@_{d:dvp}(c-goal \wedge \\ \quad \langle SpeechAct\rangle\mathbf{assertion} \wedge \\ \quad \langle Relation\rangle\mathbf{answer} \wedge \\ \quad \langle Content\rangle(e : ascription \wedge \\ \quad\quad \langle Target\rangle(t : person \wedge Bob \wedge \\ \quad\quad\quad \langle InfoStatus\rangle\mathbf{familiar}) \wedge \\ \quad\quad \langle TopIn\rangle(p : rfx \wedge RefIndex)))\end{array}$$

The content planner makes a series of decisions about the type and structure of the utterance to be produced. As it is an assertion of a property ascription, it decides to plan a sentence in indicative mood and present tense with "be" as the main verb. The reference to the target entity makes up the copula restriction, and a reference to the ascribed property is in the copula scope. This yields an expansion of the goal content:

$$\begin{array}{l}@_{e:ascription}(\mathbf{be} \wedge \\ \quad \langle Tense\rangle\mathbf{pres} \wedge \\ \quad \langle Mood\rangle\mathbf{ind} \wedge \\ \quad \langle Cop-Restr\rangle(t : entity \wedge \\ \quad\quad Bob \wedge \langle InfoStatus\rangle\mathbf{familiar}) \wedge \\ \quad \langle Subject\rangle(t : entity) \wedge \\ \quad \langle Cop-Scope\rangle(prop : m-location \wedge \\ \quad\quad \mathbf{in} \wedge \langle Anchor\rangle(p : rfx \wedge RefIndex)))\end{array}$$

The next step consists in calling the GRE algorithm to produce an RE for the entity p. In our NLP system we use a slightly modified implementation of the Incremental Algorithm [Dale and Reiter, 1995]. The context set C is determined using TAA1. Let's assume that Bob is currently in kitchen3. In our example (a =reception) the GRE algorithm hence produces the following result, which is then returned to the planner and inserted into the proto LF created so far:

$$\begin{aligned}
&@_{p:entity}(\mathbf{kitchen} \wedge \\
&\quad \langle TopOn\rangle(f : entity \wedge \\
&\quad\quad floor \wedge \langle Unique\rangle\mathbf{true} \wedge \\
&\quad\quad\quad \langle Number\rangle(n : quality \wedge 2)))
\end{aligned}$$

The planner then makes further decisions about the realization, expanding this part of the LF to the following result:

$$\begin{aligned}
&@_{p:entity}(\mathbf{kitchen} \;\wedge \\
&\quad \langle Delimitation\rangle\mathbf{unique} \;\wedge \\
&\quad \langle Num\rangle\mathbf{sg} \;\wedge\; Quantification\rangle\mathbf{specific} \;\wedge \\
&\quad \langle Modifier\rangle(o1 : m - location \;\wedge\; on \;\wedge \\
&\quad\; \langle Anchor\rangle(f : thing \;\wedge\; floor \;\wedge \\
&\quad\quad \langle Delimitation\rangle\mathbf{unique} \;\wedge \\
&\quad\quad \langle Num\rangle\mathbf{sg} \;\wedge\; \langle Quantification\rangle\mathbf{specific} \;\wedge \\
&\quad\quad \langle Modifier\rangle(t1 : number - ordinal \;\wedge\; 2))))
\end{aligned}$$

Once the planner is finished, the resulting overall LF is provided to a CCG realizer [White and Baldridge, 2003], turning it into a surface form ("Bob is in the kitchen on the second floor"). This string is synthesized to speech using the MARY TTS software [Schröder and Trouvain, 2003].

5 Conclusions and Future Work

We have presented an algorithm for context determination that can be used both for resolving and generating referring expressions in a large-scale space domain. We have presented an implementation of this approach in a dialogue system for an autonomous mobile robot.

Since there exists no suitable evaluation benchmark for situated human-robot dialogue to compare our results against, we are currently planning a user study to evaluate the performance of the TA algorithm. Another important item for future work is the exact nature of the spatial progression in situated dialogue, modeled by "moving" the referential anchor.

References

[Bateman, 1997] J. A. Bateman. Enabling technology for multilingual natural language generation: the KPML development environment. *Journal of Natural Language Engineering*, 3(1):15–55, 1997.

[Blackburn, 2000] P. Blackburn. Representation, reasoning, and relational structures: a hybrid logic manifesto. *Journal of the Interest Group in Pure Logic*, 8(3):339–365, 2000.

[Cohn and Hazarika, 2001] A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29, 2001.

[Croitoru and van Deemter, 2007] M. Croitoru and K. van Deemter. A conceptual graph approach to the generation of referring expressions. In *Proc. IJCAI-2007*, Hyderabad, India, 2007.

[Dale and Haddock, 1991] R. Dale and N. Haddock. Generating referring expressions involving relations. In *Proc. EACL-1991*, Berlin, Germany, April 1991.

[Dale and Reiter, 1995] R. Dale and E. Reiter. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[Horacek, 1997] H. Horacek. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. ACL/EACL-1997*, Madrid, Spain, 1997.

[Jacobsson *et al.*, 2008] H. Jacobsson, N. Hawes, G. J. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proc. HRI-2008*, Amsterdam, The Netherlands, 2008.

[Kelleher and Kruijff, 2006] J. Kelleher and G. J. Kruijff. Incremental generation of spatial referring expressions in situated dialogue. In *In Proc. Coling-ACL-2006*, Sydney, Australia, 2006.

[Krahmer and Theune, 2002] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA, 2002.

[Krahmer *et al.*, 2003] E. Krahmer, S. van Erk, and A. Verleg. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1), 2003.

[Kruijff *et al.*, 2009] G. J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes. Situated dialogue processing for human-robot interaction. In H. I. Christensen, G. J. Kruijff, and J. Wyatt, editors, *Cognitive Systems*. Springer, 2009. to appear.

[Kruijff, 2005] G. J. Kruijff. Context-sensitive utterance planning for CCG. In *Proc. ENLG-2005*, Aberdeen, Scotland, 2005.

[Kuipers, 1977] B. Kuipers. *Representing Knowledge of Large-scale Space*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1977.

[Lison and Kruijff, 2008] P. Lison and G. J. Kruijff. Salience-driven contextual priming of speech recognition for human-robot interaction. In *ECAI 2008*, 2008.

[Paraboni *et al.*, 2007] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June 2007.

[Schröder and Trouvain, 2003] M. Schröder and J. Trouvain. The german text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int. Journal of Speech Technology*, 6:365–377, 2003.

[Stone and Webber, 1998] M. Stone and B. Webber. Textual economy through close coupling of syntax and semantics. In *Proc. INLG-1998*, pages 178–187, Niagara-on-the-Lake, ON, Canada, 1998.

[White and Baldridge, 2003] M. White and J. Baldridge. Adapting chart realization to CCG. In *Proc. ENLG-2003*, Budapest, Hungary, 2003.

[Zender and Kruijff, 2007] H. Zender and G. J. Kruijff. Multi-layered conceptual spatial mapping for autonomous mobile robots. In *Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems*, AAAI Spring Symposium 2007, March 2007.

[Zender *et al.*, 2008] H. Zender, O. Martínez Mozos, P. Jensfelt, G. J. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June 2008.

Phrasing Questions

Geert-Jan M. Kruijff
German Research Center
for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany
gj@dfki.de

Michael Brenner
Institute for Computer Science
Albert-Ludwigs-Universität
Freiburg, Germany
brenner@informatik.uni-freiburg.de

Abstract

In a constructive learning setting, a robot builds up beliefs about the world by interacting – interacting with the world, and with other agents. Asking questions is key in such a setting. It provides a mechanism for interactively exploring possibilities, to extend and explain the robot's beliefs. The paper focuses on how to linguistically phrase questions in dialogue. How well the point of a question gets across depends on how it is put. It needs to be effective in making transparent the agent's intentions and beliefs behind raising the question, and in helping to scaffold the dialogue such that the desired answers can be obtained. The paper proposes an algorithm for deciding what to include in formulating a question. Its formulation is based on the idea of considering transparency and scaffolding as referential aspects of a question.

Introduction

Robots are slowly making their entry into ”the real world.” And it is slowly becoming an accepted fact of life that we cannot possibly provide such robots will all there is to know, out-of-the-box. So they need to learn. The point of socially guided (machine) learning (Thomaz 2006) is that some of that learning can be done effectively through social interaction with other agents in the environment.

This paper focuses on how a robot should phrase its questions, considering a social learning setting in which situated dialogue is the main interactive modality (Kruijff et al. 2006a; Jacobsson et al. 2007). The robot and a human use spoken dialogue to discuss different aspects of the environment. We consider learning to be driven by the robot's own, perceived learning needs. This requires dialogue to be mixed-initiative. Both the human and the robot can take the initiative in driving this ”show-and-tell-then-ask” dialogue. Questions play a fundamental role in such dialogues. Assuming a robot has the ability to raise issues in need of clarification or learning for any modality, (e.g. (Kruijff, Brenner, and Hawes 2008)), the problem thus becomes how to properly *phrase* a question.

Typically, a question is represented as an abstraction over the argument of a predicate. For example, assuming $?x.P(x)$ to indicate that a question regards a parameter x of some predicate $P(x)$, a question about the color of a ball could be phrased as $?x.(ball(y) \wedge has\!-\!color(y, x))$. However, more aspects need to be taken into account, for a question to be posed in such a way that the addressee is likely to understand the question and provide a suitable answer (Ginzburg 1995b).

First of all, the phrasing needs to make *transparent* how a question arises from an agent's beliefs, what beliefs – and what gaps in an agent's beliefs – it refers to. It should make clear *what a question is about*. Furthermore, there is a reason behind raising the question. The agent has a specific goal, it intends to obtain a particular kind of answer. Not just any answer will do. Raising a question also needs to set up, *scaffold*, the right context for answering it. This is the *why* of a question, pointing to how the agent would like to see the question *resolved*.

An example in (Kruijff et al. 2006b; 2007b) provides an interesting illustration.[1] The robot is capable of figuring out when it might have mistakenly classified a particular passage in the environment as a door. At the point where it realizes this, it asks, ”Is there a door here?” Unfortunately, the place where it asks this is not related to the location ”here” refers to. To anyone but a developer-acting-as-user it is not transparent what the ”here” means. This often leads to the user giving the wrong answer, namely ”yes this room has a door” rather than, ”no, there is no door between the trash bin and the table.” The way the question was phrased lacked both in transparency (location reference) and in scaffolding (specific location, not the room as such).

The paper presents an approach to generating a content representation for a question. These representations reflect what is being asked after, in reference to beliefs (aboutness, transparency) and intentions (resolvedness, scaffolding). The approach explicitly regards transparency and scaffolding as *referential qualities* of a question. This way their referential nature in the larger dialogue- and situated context can be considered. Following out that idea, the approach bases its content determination algorithm on Dale & Reiter's incremental algorithm for generating referring expressions (Dale and Reiter 1995), in combination with algo-

[1] See also the video at the CoSy website's Explorer page, at http://cosy.dfki.de/www/media/explorer.y2.html.

rithms for referential context determination (Zender, Kruijff, and Kruijff-Korbayová 2009; Paraboni, van Deemter, and Masthoff 2007).

Central to the approach is establishing the information pertaining to the question. A description logic-like formalism is used to represent such information, as a conceptual structure in which propositions have ontological sorts and unique indices, and can be related through named relations. A question can then be represented as a structure in which we are querying one or more aspects of such a representation (Ginzburg 1995b; Kruijff, Brenner, and Hawes 2008). The formalism allows everything to be queried: relations, propositions, sorts. Around the formulation of a question we construct a nucleus, comprising the situation (the "facts") and the beliefs that have led up to the question, the question itself, and the goal content which would resolve the question. The question nucleus integrates Ginzburg's notions of aboutness, and (potential) resolvedness.

Based on the question nucleus, the algorithm starts by determining to what extend the different aspects are covered by the (dialogue) common ground between the robot and the human. For this, contextual references are resolved in a dialogue context model (Kruijff et al. 2007a), and it is established how these can be related to inferences over domain knowledge and instances (Kruijff et al. 2007b). The question nucleus is extended with these connections – or rather, with indications of the information structure or informativity of individual content – so that it includes an explicit notion of what is shared, and what is privately held information (cf. (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999)).

The algorithm next decides what aspects of a question nucleus to include in the content for phrasing the question. For each aspect of the nucleus (facts, beliefs, question, goals) the algorithm uses the informativity of the aspect's content, in conjunction with similarly related but contrasting content in the dialogue context model, to determine whether to include it. Essentially, new or contrastive content will be considered, whereas salient "old" information will not. The form in which the content will be included is determined by content-specific algorithms for generating referring expressions (e.g. (Kelleher and Kruijff 2006; Zender, Kruijff, and Kruijff-Korbayová 2009)). The decisions to include particular content can be weighted according to a comprehensibility ranking as e.g. in (Krahmer, van Erk, and Verleg 2003).

The contributions the approach aims for are, briefly, as follows. Purver and Ginzburg develop an account for generating questions in a dialogue context (Purver, Ginzburg, and Healey 2003; Purver 2004). Their focus was, however, on clarification for the purpose of dialogue grounding. A similar observation can be made for recent work in HRI (Li, Wrede, and Sagerer 2006), We are more interested in formulating questions regarding issues in building up situation awareness, including the acquisition of new ways of understanding situations (cf. also (Kruijff, Brenner, and Hawes 2008)). In issue-based (or information state-based) dialogue systems (Larsson 2002), the problem of how to phrase a question is greatly simplified because the task domain is fixed. There is little need for paying attention to transparency or scaffolding, as it can be assumed the user understands the task domain.

An overview of the paper is as follows.The paper starts with a discussion of basic issues in modeling questions and their semantics, based on (Ginzburg 1995b). Then the approach is presented. The approach starts from the assumption that a question is a dialogue, not just a single utterance. Discussed is how the content plan for such a question dialogue can be determined, providing definitions, representation, and algorithms. The paper ends with a discussion of how the approach could be integrated, evaluated, and points for further research.

Background

What is a question? Ginzburg (1995b) discusses a variety of linguistic approaches. All of them aim to provide an invariant characterization of the semantics of a question. Broadly, they have proposed the following aspects as crucial to that definition.

First, several approaches propose to see a question as an n-ary relation. The relation puts together the question with one or more contributions pertaining to answering it. The point here is to take into account the fact that a question can be discussed over several turns in a dialogue. Second, there is a sense of *aboutness* to a question. Each question can be associated with a collection of propositions, which are –intuitively– related to the question. And, finally, each question can be considered to be associated with a (possibly complex) proposition which provides an *exhaustive answer*. In other words, an exhaustive answer resolves the question.

Ginzburg suggests that all these aspects together make up a characterization of a question – not just one of them, as most approaches suggest. Furthermore, these aspects are to be understood as being *relative*. What a question is about, and how it can be resolved, should be understood relative to an agent's *goal* and *belief/knowledge state* (cf. also (Ginzburg 1995a)). The following example illustrates this.

(1) Context: a robot drives around campus, and is about to enter the DFKI building.

a. Janitor: Do you know where you are?
Robot: DFKI.

b. Janitor believes the robot knows where it is.

(2) Context: a robot drives around the DFKI building, to get a cup of coffee.

a. Janitor: Do you know where you are?
Robot: DFKI.

b. The janitor is not convinced the robot really knows where it is.

What counts as an answer to a question may thus vary across contexts. What a question is thus cannot be reduced to an analysis of just what counts as its answers. Instead, Ginzburg starts with setting up an ontology in which questions, propositions and facts are considered as equal citizens. This makes it possible to consider a question *in relation to* possible answers for it. The ontology is defined using situation theoretic constructs, which we will adopt throughout this paper. (All definitions as per (Ginzburg 1995a; 1995b).)

Definition 1 (SOA, Situation, Fact)**.** A SOA (State Of Affairs) describes possible ways an actual situation might be. SOAs are either *basic*, or built up from basic ones using algebraic operations. A *basic SOA* is an atomic possibility, written as $\langle R, f : i \rangle$ with R a relation, f a mapping assigning entities to the argument roles of R, and i is a polarity i.e. $i \in \{+, -\}$. A situation s supports the factuality of a SOA σ iff $s \models \sigma$. The SOA σ is then considered a *fact* in s. To enable complex SOAs, SOAs can be structured as a Heyting algebra under a partial order '$\rightarrow$', which is closed under arbitrary meets ($\bigwedge$) and joins ($\bigvee$). Situations and SOAs together form a SOA-algebra:

1. If $s \models \sigma$ and $\sigma \rightarrow \tau$ then $\sigma \models \tau$
2. $s \not\models 0$, $s \models 1$ (FALSE,TRUE)
3. If Σ is any finite set of SOAs, then $s \models \bigwedge \Sigma$ iff $s \models \sigma$ for each $\sigma \in \Sigma$
4. If Σ is any finite set of SOAs, then $s \models \bigvee \Sigma$ iff $s \models \sigma$ for at least one $\sigma \in \Sigma$

Finally, an application operator is defined, to allow for variable assignment (and reduction):
$\lambda x.\langle R, a : b, c : x : + \rangle [x \mapsto d] = \langle R, a : b, c : d : + \rangle$ □

Using Definition 1, we can now consider a proposition to be an assertion about the truth of a possibility relative to a situation.

Definition 2 (Proposition)**.** A proposition p is a relational entity, asserting a truth regarding a SOA τ in a particular situation s: $p = (s : \tau)$. A proposition $p = (s : \tau)$ is TRUE iff τ is a *fact* of s, denoted as $s \models \tau$. □

Before defining what a question is, the notions of *resolvedness* and *aboutness* need to be defined. Resolvedness, or rather the broader concept of *potentially resolving* a question, is defined as follows. The definition distinguishes whether a (possibly complex) fact resolves a question depending on whether the question is *polar*, asking for the truth of an assertion (e.g. "Is the ball red?"), or *factive*, asking after a value (e.g. "What color is the ball?").

Definition 3 (Resolvedness conditions)**.** A SOA τ *potentially resolves* a question q if either

1. τ positively-resolves q (for 'polarity p': any information that *entails* p; for a factive question: any information that entails that the extension of the queried predicate is non-empty)
2. τ negatively-resolves q (for 'polarity p': any information that *entails* $\neg p$; for a factive question: any information that entails that the extension of the queried predicate is empty)

□

We will leave the notion of *aboutness* for the moment. Essentially, Ginzburg (1995a; 1995b) defines this as a collection of SOAs which can be associated with the content of a question q, with a SOA being about q if it subsumes the fact that q is either positively or negatively resolved. (For subsumption, recall Definition 1.)

Ginzburg's definition of what a question is then works out as follows.

Definition 4 (Question)**.** A question is an entity $(s?\mu)$ constructed from a situation s and an n-ary abstract SOA $\mu = \lambda x_1, ..., x_n \sigma(x_1, ..., x_n)$ $(n \geq 0)$:

1. μ constitutes an underspecified SOA from which the class of SOAs that are *about* q can be characterized.
2. Those SOAs which are facts of s and informationally subsume a level determined by μ constitute a class of SOAs that *potentially* resolve q.

□

The definition includes references to the relational character of a question (the abstract), and the notions of aboutness (intuitively, the space within which we are looking for an answer) and of resolvedness (the space of possible answers we are looking for, one of which will -hopefully- establish itself as fact). Finally, we already indicated above that resolvedness is an agent-relative notion. Ginzburg suggests to do so using Definition 3 as follows.

Definition 5 (Agent-relative resolvedness)**.** A fact τ *resolves* a question $(s?\mu)$ relative to a mental situation ms iff

1. Semantic condition: τ is a fact of s that potentially resolves μ
2. Agent relativisation: $\tau \implies_{ms} Goal - content(ms)$, i.e. τ entails the goal represented in the mental situation ms relative to the inferential capabilities encoded in ms.

□

Approach

The previous section presented a formal (but relatively abstract) notion of what a question is. It made clear that a question is more than a predicate with an open variable, or (alternatively) just another way of characterizing a set of propositions that would serve as exhaustive answer. Instead, a question is a relational structure, tying into a larger context. For one, this "context" provides a set of beliefs (SOAs, in Ginzburg's terms), a background within which potential answers are sought. An agent's goals help motivate to focus which beliefs are associated with the question. Another point about this "context" is that a question isn't just a single utterance, or just forming a unit with an utterance that answers it. There is a dialogue context in which this question is phrased. The question itself, and whatever utterances contribute to help clarify, refine and answer that question, may (though need not) refer to content already established in that context.

Phrasing a question, in other words, means we need to provide the possibility for such contextual factors to influence how the content of a question is determined. Once the agent has determined that it needs to raise a question, and about what (e.g. cf. (Kruijff, Brenner, and Hawes 2008) for questions in situated forms of learning), it needs to establish how best to communicate the question. In this paper, we suggest to do this as follows. We will begin by further explication of the notion of question, using a structure we term the *question nucleus*. The question nucleus captures more explicitly the relation between beliefs and intentions that are active in a current context, and how they determine the space of possible answers (or complexes of those). Then, we sketch several algorithms. The first group of algorithms concern *context determination*. Intuitively, these algorithms determine what beliefs and potential answers form the relevant background for the question. The background specifies what can be assumed to be known, (and can thus be referred to or even silently assumed), both in terms of content and intentions in the the dialogue- and situated context. How a question is to be phrased relies on what it needs to explicate relative to that background, to effectively communicate it. This is then finally done by the *content determination* algorithm. The result of this algorithm is a logical form, expressed in a (decidable) description logic. The logical form specifies the core content for the question, which a content planner subsequently can turn into one or more fully-fledged utterances.

The following definition defines more precisely what we mean by a logical form, based on (Blackburn 2000; Baldridge and Kruijff 2002). We will use the same formalism to describe SOAs (cf. Definition 1).

Definition 6 (Logical forms)**.** A logical form is a formula ϕ built up using a sorted description logic. For a set of propositions $PROP = \{p, ...\}$, an inventory of ontological sorts $SORT = \{s, ...\}$, and a set of modal relations $MOD = \{R, ...\}$, $\phi = p \mid i : s \mid \psi \wedge \psi' \mid \langle R \rangle \psi \mid @_{i:s}\psi$. The construction $i : s$ identifies a nominal (or index) with ontological sort s. The at-operator construction $@_{i:s}\psi$ specifies that a formula ψ holds at a possible world uniquely referred to by i, and which has ontological sort s. □

A standard Kripke-style model-based semantics can be defined for this language (Blackburn 2000). Intuitively, this language makes it possible to build up relational structures, in which propositions can be assigned ontological sorts, and referred to by using i as indices. For example, $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (c1 : color \wedge \textbf{red})$ means we have a "ball" entity, which we can uniquely refer to as $b1$, and which has a (referable) color property. (An alternative, equal way of viewing this formula is as a conjunction of elementary predications: $@_{b1:entity}\textbf{ball} \wedge @_{b1:entity}\langle Property \rangle c1 : color \wedge @_{c1:color}\textbf{red}$.)

Question nucleus

We start by defining the notion of *question nucleus*. The function of a question nucleus is twofold. First, it should capture the question's background in terms of associated beliefs and intentions, and what space of expected answers these give rise to. An expected answer is naturally only as specific (or unspecific) as is inferable on the basis of what the agent knows.

Definition 7 (Expected answer)**.** An expected answer a for a question q is a proposition $a = (s : \tau)$, with τ potentially resolving q as per Definition 3. τ is a logical formula (Definition 6) which can be underspecified, both regarding the employed ontological sorts, and arguments. □

Effectively, assuming that the agent has a collection of ontologies which provide a subsumption structure ($a \sqsupset b$ meaning a subsumes b, i.e. b is more specific), an expected answer can be said to define a "level" of specifity (Definition 4) according to subsumption. Following up on the ball example, assume the agent has an ontology which defines $material - property \sqsupset \{color, shape\}$. An expected answer to a question, what particular shape the ball has, would take the form $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (s1 : shape))$. All the proposition specifies is that there is an identifiable shape. If the question would be about any, or some unknown, property of the ball, an expected answer could be phrased as $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (m1 : material - property))$. Using the available ontological structure, and relational structure between formulas, we can formulate expected answers at any level of specifity without requiring the agent to already know the answer (cf. also (Kruijff, Brenner, and Hawes 2008)).

Definition 8 (Question nucleus)**.** A *question nucleus* is a structure $qNucleus = \{r, BL, XP, AS\}$ with:

1. A referent r relative to which the question q (part of XP) is phrased.
2. BL (*Beliefs*) is a set of private and shared beliefs, about agent intentions and facts in the current context (cf. (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999)).
3. XP (*Execution Plan*) is a continual plan with an execution record (Brenner and Nebel 2008) for resolving a question $q = (s?\mu)$.
4. AS (*Answer Structure*) is a finite $\sqsupset$-structure over propositions $p_1, ...$ which potentially resolve q, and which are implied by BL.

The beliefs BL specify what the agent knows about r, what the agent presumes to be shared knowledge about r, and what the agent presumes other agents could know about r. BL is based on the dialogue leading up to the question, any previous actions involving r, and a domain model of agent competences (Brenner and Kruijff-Korbayová 2008). XP makes explicit that phrasing a question constitutes a dialogue, with an associated plan for communicating the question and a record for how far the question has been fully answered. This record maintains which aspects (elementary predications) of the question are still open ("under discussion," similar to the Question-Under-Discussion construct of (Ginzburg 1995b)). The AS is a set of propositions, relating those propositions to the aspect(s) of the question they would potentially resolve (and thus to the execution record in XP). AS is based on propositions implied by BL (relative to r, q) and is $\sqsupset$-structured according to ontological structure. □

Contextually determining aboutness

Asking a question starts with the agent having determined what it is it needs to know about some referent r, e.g. an area in the environment, an object – or, more specifically, relations or properties. (To allow for group referents, we will consider r to be a *set*.) Next the question nucleus is built up, starting with the beliefs about the question, BL.

We adopt the approach to belief modeling described in (Brenner and Kruijff-Korbayová 2008). Beliefs are formulated as relational structures with *multi-valued state variables* (MVSVs). These state variables are used for several purposes. First, they can indicate domain values, as illustrated by the sorted indices in the examples above. The color $c1$ would be a Property-type state variable of the entity $b1$, and could take domain values in the range of that ontological sort. Important is that the absence of a value for an MVSV is interpreted as *ignorance*, not as falsehood: $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(s1 : shape))$ means the agent does not know what shape the ball has, not that it has no shape (as per a closed-world assumption). In a similar way, state variables are used for expressing *private beliefs*, and mutual or *shared beliefs* (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999). A private belief of agent a_1 about content ϕ is expressed as $(K\{a_1\}\phi)$ whereas a mutual belief, held by several agents, is expressed as $(K\{a_1, a_2, ...\}\phi)$. Secondly, MSVSs can be quantified over, for example using the ? to express a question: $?s1.@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(s1 : shape))$ represents a question regarding the shape of the referent $b1$.

As an agent perceives the environment, we assume it builds up beliefs about the instances it perceives, and what relations can be observed or inferred to hold between them. For example, see (Brenner et al. 2007) for a robot manipulating objects in a local visual scene, or (Kruijff et al. 2007b) for a robot exploring an indoor environment. Furthermore, we assume that the agent's planning domains include models of agent capabilities – what another agent is capable of doing, including talking (and answering questions!) about particular aspects of the environment (Brenner and Kruijff-Korbayová 2008). Finally, if the agent has been engaged in a dialogue with another agent, and discussed the referent-in-question r before, we assume that the (agreed-upon) content discussed so far constitutes shared beliefs, held by all agents involved.

Algorithm 1 : Determine(BL) (*sketch*)

Require: BELs is a set of private and mutual beliefs the agent holds, (including beliefs about capabilities); r is the referent (set) in question

```
BL = $\emptyset$
for b $\in$ BELs do
    if b includes a MVSV $m \in r$ then
        BL = BL $\cup$ b
    end if
end for
return BL
```

Algorithm 1 sketches the basis of the algorithm for establishing BL. Those beliefs are gathered which refer explicitly to the referent the question is about. Note that BL may end up being empty. This means that r has not been talked about, nor does the agent know whether another agent could actually offer it an answer to what it would like to know more about.

Contextually determining resolvedness

The beliefs BL about the referent in question r state what the agent already believes about r (privately, or shared), and what it believes about another agent's capabilities. Next, these beliefs need to be structured such that potentially resolving answers can be derived. We assume that we can make use of the ontological sorts, and the structuring over these sorts provided by domain ontologies, to organize beliefs. The organization we are after first of all relates a belief to a potentially resolving answer, by combining it (inferentially) with the ?-quantified, ontologically sorted MVSVs in the question to yields a partially or completely reduced logical form (Definition 1). Secondly, the organization relates beliefs by (sortal) subsumption over the potentially resolving answers they generate.

For example, consider a question about the color of a ball: $?c1.@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(c1 : color))$. Let us assume the robot holds several beliefs with regard to $b1$, and the variable $c1$. A robot learning more about visual properties of objects through interaction with a human tutor (Jacobsson et al. 2007) typically holds at least beliefs about what the tutor is capable of telling it. Thus, assume the robot believes the tutor can tell it about material properties, colors, and shapes. Using tell-val (*tell value* action) we can model these beliefs as $(K\ \{a_1\}\ tell - val(a_2, m : material - property)$, $(K\ \{a_1\}\ tell - val(a_2, c : color)$. The variables m, b are existentially bound in these beliefs. Using the inference that $material - property \ \sqsupset \ color$ and introducing bound variables m', c' for m and c respectively, the beliefs can be combined with the question to yield the potentially resolving propositions $c' : color, m' : material - property$. Furthermore, subsumption yields $m' : material - property \sqsupset c' : color$. Thus, by combining the beliefs with what the agent already knows, it can expect to know something it doesn't yet know by asking a question. And by making use of the way its knowledge is ontologically structured, it can determine how precise that answer is likely to be.

Algorithm 2 provides a first sketch of the algorithm for establishing AS. (In the current version, propositional content and additional relational structure pertaining to m in the context of b is not yet included into AS.)

Content determination

Finally, once the beliefs about q and the potentially resolving answers for q have been established, we can turn to determining the exact content for communicating q. The purpose of content determination is to establish what, how much, should be communicated for the agent to get an appropriate answer – how much content it needs to communicate to ensure proper scaffolding and transparency. For example,

Algorithm 2 : Determine(AS) (*sketch*)

```
Require: BL is a set of beliefs relative to $r$, $q$ is a question
  about $r$, and ONT is a collection of ontologies supporting
  subsumption inferences on sorts used in $BL$ and $q$.

  AS = $\emptyset$ (empty subsumption )
  for b $\in$ BLs do
    $\phi$ = $\top$
    for MVSV $m \in r$ existentially bound in $b$ do
      introduce a bound variable $m'$
      $\phi = \phi \wedge m' : sort(MVSV)$
    end for
    AS = AS $\sqcup$ $\phi$, under $\sqsupset$
  end for

  return AS
```

consider again the question about the color of the ball. How the question should be phrased, depends on whether e.g. the ball has already been talked about, what goals are involved (are we learning how this ball looks like, or how objects roll?), etc. Example 3 provides some illustrations.

(3) Asking about the color of a single ball on a table ...

a. If the robot is not sure whether the other agent knows about colors:
"Could you tell me about the color of this ball?"

b. If the robot believes the other agent knows about colors:
" Could you tell me what color this ball is?"

c. If the robot is not sure whether asking about color is relevant to the current goal:
"I would like to know more about the color of this ball. Could you tell me what it is?"

d. If the ball is under discussion, and asking for color is relevant:
"What's the color?"

Example 3 particularly illustrates how scaffolding and transparency come into play. We connect these terms explicitly to the question nucleus. We see scaffolding primarily as appropriately embedding a question into an intentional setting, relating to AS and the extent to which available beliefs lead to specific (potentially resolving) answers. Transparency relates to the referential setting of the question nucleus, relating r to BL in the sense of what the agent can already assume to be mutually known about the referent under discussion. Planning the question as a dialogue, then, means determining relevant beliefs, and the information status of relevant content. Relevant beliefs are those which are associated with maximally specific, potentially resolving answer(s). A distinction needs to be made between private and mutual believes, particularly as beliefs about competences are first and foremost private beliefs. Furthermore, it should be determined whether these beliefs fit into the current intentional context. (For the purposes of the current paper, we will consider learning goals only, and consider them to specify what ontological sorts the agent is trying to learn.) Information status regards whether content, pertaining to r, can be assumed to be mutually known – most notably, whether r is mutually known (i.e. mutually identifiable in context).

Algorithm 3 : Content determination (*sketch*)

```
Require: BL is a set of beliefs relative to $r$, $q$ is a question
  about $r$, ONT is a collection of ontologies supporting sub-
  sumption inferences on sorts used in $BL$ and $q$, $AS$ is a
  structure over potentially resolving answers

  RelBL = $\emptyset$
  for a $\in$ AS do
    if a is maximally specific, i.e. there is no a' s.t. a $\sqsupset$
    a' then
      RelBL = RelBL $\cup$ { b }, for b yielding a
    end if
  end for
  MutualRelBL = mutual beliefs in RelBL
  ScaffoldingBL = $\emptyset$
  TransparencyBL = $\emptyset$
  for MVSV $m$ in $q$ do
    if there is a $b \in$ MutualRelBL associated to $m$ then
      TransparencyBL = TransparencyBL $\cup$ { b }
    else
      ScaffoldingBL = ScaffoldingBL $\cup$ { be-
      liefs associated to most specific answers for $m$ }
    end if
  end for
  return ScaffoldingBL, TransparencyBL
```

Algorithm 3 first determines what beliefs are relevant to achieve a maximally specific answer, and which of these beliefs are mutual. How much scaffolding needs to be done depends on whether these mutual beliefs imply all potentially resolving answers to the questioned MVSVs in r. If not, the algorithm backs off by constructing a belief set which needs to be communicated for appropriate scaffolding. The basis for transparency is formed by the mutual beliefs about r.

On the basis of these sets of beliefs, and q itself, the communication of q can be planned. We do not provide an in-depth discussion of dialogue- and content-planning here, for space (and time) reasons. We refer the interested reader to (Brenner and Kruijff-Korbayová 2008; Kruijff et al. 2009). In brief, beliefs in the scaffolding set are specified as assertions (Brenner and Nebel 2008). The plan for communicating the question starts by verifying these assertions, and then raises the question itself. It is a matter for content fusion whether such verification can be done in conjunction with the question itself (Example 3, a–b) or as preceding utterances (Example 3, c). For the realization of the question, the transparency beliefs are used to determine information status. Content planning then turns information status into decisions about how to refer to r and the asked-after properties – e.g. using pronominal reference (Example 3, c) or even omitting explicit reference, by eliding any mention of r (Example 3, d).

Conclusions

The approach presented in this paper is still under development. The key technologies it is based on (planning, motivation, dialogue processing, and ontological inferencing) are already available in the system architecture the approach will be integrated into. We will describe the full integration, with working examples, in a full version of this paper. We will then also consider how this approach can be applied in related settings, such as performance requests.

We are currently considering various alternative ways to evaluate the approach. User experiments are just one option here. The problem is that an approach as presented here, and the overall architecture it will be integrated into, present a large parameter space. Consequently, it is difficult to ensure a controlled setting for a user experiment – and, only a very limited part of the parameter space can be effectively explored. An alternative way we are therefore currently considering is to use techniques from language evolution. In simulations we would like to explore what the effects of different parameter settings would be on how agents are able to communicate, and what this consequently means for measurable parameters such as learning performance. Examples of such experiments can be found in (Ginzburg and Macura 2006).

There remain for the moment plenty of open issues to be investigated further – this paper really only provides a first description of the approach we are developing. It does aim to make clear how notions such as scaffolding and transparency can be folded into a characterization of how a system can phrase a question – seeing a question, in fact, as a subdialogue to be planned, not just a single utterance paired with a possible answer. Basic issues remain in the construction of the various belief sets, and the associated structures over potentially resolving answers. Although an "unweighted" approach as followed here will work for most simple scenarios, it remains to be seen whether associating *costs* with beliefs (and assuming them, in a plan for communicating a dialogue) could provide a more adaptive, scalable approach in the long run. Furthermore, the current formulation of the construction of the answer structure AS (Algorithm 2) does not cover polar questions (though this is an easy extension).

Acknowledgments

This work was supported by the EU FP7 IST Project "CogX" (FP7-IST-215181).

References

Baldridge, J., and Kruijff, G. 2002. Coupling CCG and hybrid logic dependency semantics. In *Proc. ACL 2002*, 319–326.

Blackburn, P. 2000. Representation, reasoning, and relational structures: a hybrid logic manifesto. *Logic Journal of the IGPL* 8(3):339–625.

Brenner, M., and Kruijff-Korbayová, I. 2008. A continual multiagent planning approach to situated dialogue. In *Proceedings of the LONDIAL (The 12th SEMDIAL Workshop on Semantics and Pragmatics of Dialogue)*.

Brenner, M., and Nebel, B. 2008. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*.

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

Dale, R., and Reiter, E. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Ginzburg, J., and Macura, Z. 2006. Lexical acquisition with and without metacommunication. In Lyon, C.; Nehaniv, C.; and Cangelosi, A., eds., *The Emergence of Communication and Language*. Springer Verlag. 287–301.

Ginzburg, J. 1995a. Resolving questions, I. *Linguistics and Philosophy* 18(5):459–527.

Ginzburg, J. 1995b. The semantics of interrogatives. In Lappin, S., ed., *Handbook of Contemporary Semantic Theory*. Blackwell.

Grosz, B., and Kraus, S. 1999. The evolution of shared plans. In Rao, A., and Wooldridge, M., eds., *Foundations and Theories of Rational Agency*. Springer. 227–262.

Jacobsson, H.; Hawes, N.; Skocaj, D.; and Kruijff, G. 2007. Interactive learning and cross-modal binding – a combined approach. In *Language and Robots: Proceedings of the Symposium*, 1pp–1pp.

Kelleher, J., and Kruijff, G. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 1041–1048.

Krahmer, E.; van Erk, S.; and Verleg, A. 2003. Graph-based generation of referring expressions. *Computational Linguistics* 29(1):53–72.

Kruijff, G.; Kelleher, J.; Berginc, G.; and Leonardis, A. 2006a. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*.

Kruijff, G.; Zender, H.; Jensfelt, P.; and Christensen, H. 2006b. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*.

Kruijff, G.; Lison, P.; Benjamin, T.; Jacobsson, H.; and Hawes, N. 2007a. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*.

Kruijff, G.; Zender, H.; Jensfelt, P.; and Christensen, H. 2007b. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems* 4(2).

Kruijff, G.; Lison, P.; Benjamin, T.; Jacobsson, H.; Zender, H.; and Kruijff-Korbayová, I. 2009. Situated dialogue processing for human-robot interaction. In Christensen, H.; Kruijff, G.; and

Wyatt, J., eds., *Cognitive Systems*. Available at http://www.cognitivesystems.org/cosybook.

Kruijff, G.; Brenner, M.; and Hawes, N. 2008. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*.

Larsson, S. 2002. *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden.

Li, S.; Wrede, B.; and Sagerer, G. 2006. A computational model of multi-modal grounding. In *Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, 153–160.

Lochbaum, K.; Grosz, B.; and Sidner, C. 1999. Discourse structure and intention recognition. In Dale, R.; Moisl, H.; ; and Somers, H., eds., *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. New York: Marcel Dekker.

Paraboni, I.; van Deemter, K.; and Masthoff, J. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics* 33(2):229–254.

Purver, M.; Ginzburg, J.; and Healey, P. 2003. On the means for clarification in dialogue. In Smith, R., and van Kuppevelt, J., eds., *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*. Kluwer Academic Publishers. 235–255.

Purver, M. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. Dissertation, King's College, University of London.

Thomaz, A. L. 2006. *Socially Guided Machine Learning*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Zender, H.; Kruijff, G.; and Kruijff-Korbayová, I. 2009. A situated context model for resolution and generation of referring expressions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 126–129.

Efficient Parsing of Spoken Inputs for Human-Robot Interaction

Pierre Lison and Geert-Jan M. Kruijff

Abstract— **The use of deep parsers in spoken dialogue systems is usually subject to strong performance requirements. This is particularly the case in human-robot interaction, where the computing resources are limited and must be shared by many components in parallel. A real-time dialogue system must be capable of responding quickly to any given utterance, even in the presence of noisy, ambiguous or distorted input. The parser must therefore ensure that the number of analyses remains bounded at every processing step.**

The paper presents a practical approach to address this issue in the context of deep parsers designed for spoken dialogue. The approach is based on a word lattice parser combined with a statistical model for parse selection. Each word lattice is parsed incrementally, word by word, and a discriminative model is applied at each incremental step to prune the set of resulting partial analyses. The model incorporates a wide range of linguistic and contextual features and can be trained with a simple perceptron. The approach is fully implemented as part of a spoken dialogue system for human-robot interaction. Evaluation results on a Wizard-of-Oz test suite demonstrate significant improvements in parsing time.

I. Introduction

Developing robust and efficient parsers for spoken dialogue is a difficult and demanding enterprise. This is due to several interconnected reasons.

The first reason is the pervasiveness of *speech recognition errors* in natural (i.e. noisy) environments, especially for open, non-trivial discourse domains. Automatic speech recognition (ASR) is indeed a highly error-prone task, and parsers designed to process spoken input must therefore find ways to accomodate the various ASR errors that may (and will) arise. This problem is particularly acute for robots operating in real-world noisy environments and deal with utterances pertaining to complex, open-ended domains.

Next to speech recognition, the second issue we need to address is the *relaxed grammaticality* of spoken language. Dialogue utterances are often incomplete or ungrammatical, and may contain numerous disfluencies like fillers (err, uh, mm), repetitions, self-corrections, etc. Rather than getting crisp-and-clear commands such as "*Put the red ball inside the box!*", we are more likely to hear utterances such as: "*right, now, could you, uh, put the red ball, yeah, inside the ba/ box!*". This is natural behaviour in human-human interaction [1] and can also be observed in several domain-specific corpora for human-robot interaction [2]. Spoken dialogue parsers should therefore be made robust to such ill-formed utterances.

Finally, the vast majority of spoken dialogue systems are designed to operate in *real-time*. This has two important consequences. First, the parser should not wait for the utterance to be complete to start processing it – instead, the set of possible semantic interpretations should be gradually built and extended as the utterance unfolds. Second, each incremental parsing step should operate under strict time constraints. The main obstacle here is the high level of ambiguity arising in natural language, which can lead to a combinatorial explosion in the number of possible readings.

The remaining of this paper is devoted to addressing this last issue, building on an integrated approach to situated spoken dialogue processing previously outlined in [3], [4]. The approach we present here is similar to [5], with some notable differences concerning the parser (our parser being specifically tailored for robust spoken dialogue processing), and the features included in the discriminative model.

An overview of the paper is as follows. We first describe in Section II the cognitive architecture in which our system has been integrated. We then discuss the approach in detail in Section III. Finally, we present in Section IV the quantitative evaluations on a WOZ test suite, and conclude.

II. Architecture

The approach we present in this paper is fully implemented and integrated into a cognitive architecture for autonomous robots. A recent description of the architecture is provided in [6], [7]. It is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks. Figure 1 illustrates the architecture schema for the communication subsystem, limited to the comprehension side.

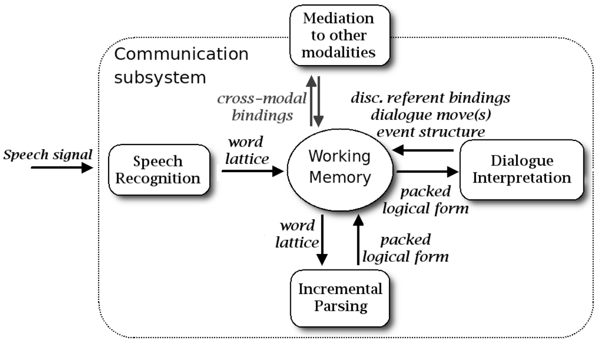


Fig. 1. Architecture schema of the communication subsystem (limited to the comprehension part).

This work was supported by the EU FP7 ICT Integrated Project "CogX" (FP7-ICT- 215181).

Pierre Lison and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany {pierre.lison},{gj} @ dfki.de

Starting with ASR, we process the audio signal to establish a *word lattice* containing statistically ranked hypotheses about word sequences. Subsequently, parsing constructs grammatical analyses for the given (partial) word lattice. A grammatical analysis constructs both a syntactic analysis of the utterance, and a representation of its meaning. The analysis is based on an incremental chart parser[1] for Combinatory Categorial Grammar [8]. These meaning representations are ontologically richly sorted, relational structures, formulated in a (propositional) description logic – more precisely in the HLDS formalism [9]. The parser itself is based on a variant of the CKY algorithm [10].

Once all the possible (partial) parses for a given (partial) utterance are computed, they are filtered in order to retain only the most likely interpretation(s). This ensures that the number of parses at each incremental step remains bounded and avoid a combinatorial explosion of the search space. The task of selecting the most likely parse(s) among a set of possible ones is called *parse selection*. We describe it in detail in the next section.

At the level of dialogue interpretation, the logical forms are then resolved against a dialogue model to establish co-reference and dialogue moves.

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* accross modalities [11].

A. Context-sensitivity

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of the utterance as it unfolds [12]. During utterance comprehension, humans combine linguistic information with scene understanding and "world knowledge" to select the most likely interpretation.

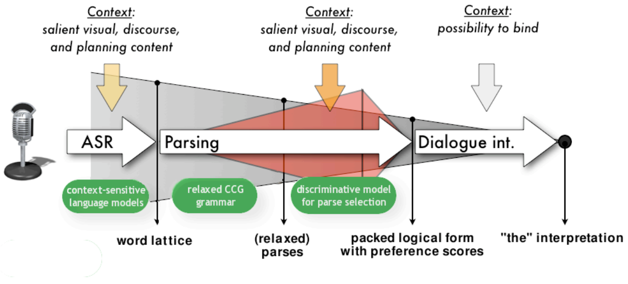

Context:
salient visual, discourse, and planning content
Context:
salient visual, discourse, and planning content
Context:
possibility to bind
ASR
Parsing
Dialogue int.
context-sensitive language models
relaxed CCG grammar
discriminative model for parse selection
word lattice
(relaxed) parses
packed logical form with preference scores
"the" interpretation

Fig. 2. Context-sensitivity in processing situated dialogue understanding

Several approaches in situated dialogue for human-robot interaction have made similar observations [13], [14], [15], [7]: A robot's understanding can be improved by relating utterances to the situated context. By incorporating contextual information into our model, our approach to robust processing of spoken dialogue seeks to exploit this important insight. At each processing step (speech recognition, word lattice parsing, dialogue-level interpretation and cross-modal binding), contextual information is used to prime the utterance comprehension, as shown in the Figure 2.

III. Approach

As we just explained, the parse selection module is responsible for selecting at each incremental step a subset of "good" parses. Once the selection is made, the best analyses are kept in the parse chart, while the others are discarded and pruned from the chart.

A. The parse selection task

To achieve this selection, we need a mechanism to discriminate among the possible parses. This is done via a (discriminative) statistical model covering a large number of features.

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

1) A function $\mathbf{GEN}(x)$ which enumerates all possible parses for an input x. In our case, the function represents the admissibles parses 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 help us discriminate between the various parses.
3) A parameter vector $\mathbf{w} \in \Re^d$.

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

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

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

The task of parse selection is an example of a *structured classification problem*, which is the problem of predicting an output y from an input x, where the output y has a rich internal structure. In the specific case of parse selection, x is a word lattice, and y a logical form.

[1] Built using the OpenCCG API: http://openccg.sf.net

B. Training data

In order to estimate the parameters $\mathbf{w}$, we need a set of training examples. Unfortunately, no corpus of situated dialogue adapted to our task domain is available to this day, let alone semantically annotated. The collection of in-domain data via Wizard of Oz experiments being a very costly and time-consuming process, we followed the approach advocated in [16] and *generated* a corpus from a hand-written task grammar.

To this end, we first collected a small set of WoZ data, totalling about a thousand utterances related to a simple scenario of object manipulation and visual learning. This set is too small to be directly used as a corpus for statistical training, but sufficient to capture the most frequent linguistic constructions in this particular context. Based on it, we designed a domain-specific context-free grammar covering most of the utterances. Each rule is associated to a semantic HLDS representation. Weights are automatically assigned to each grammar rule by parsing our corpus, hence leading to a small *stochastic context-free grammar* augmented with semantic information.

Once the grammar is specified, it is randomly traversed a large number of times, resulting in a larger set (about 25.000) of utterances along with their semantic representations. Since we are interested in handling errors arising from speech recognition, we also need to "simulate" the most frequent recognition errors. To this end, we *synthesise* each string generated by the domain-specific grammar, using a text-to-speech engine[2], feed the audio stream to the speech recogniser, and retrieve the recognition result.

Via this technique, we are able to easily collect a large amount of training data. Because of its relatively artificial character, the quality of such training data is naturally lower than what could be obtained with a genuine corpus. But, as the experimental results will show, it remains sufficient to train the perceptron for the parse selection task, and achieve significant improvements in accuracy and robustness. In a near future, we plan to progressively replace this generated training data by a real spoken dialogue corpus adapted to our task domain.

C. Perceptron learning

The algorithm we use to estimate the parameters $\mathbf{w}$ using the training data is a **perceptron**. The algorithm is fully online - it visits each example in turn, in an incremental fashion, and updates $\mathbf{w}$ if necessary. Albeit simple, the algorithm has proven to be very efficient and accurate for the task of parse selection [5], [17].

The pseudo-code for the online learning algorithm is detailed in [**Algorithm 1**].

It works as follows: the parameters $\mathbf{w}$ are first initialised to some arbitrary values. Then, for each pair (x_i, z_i) in the training set, the algorithm searchs for the parse y' with the highest score according to the current model. If this parse happens to match the best parse which generates z_i (which we shall denote y^*), we move to the next example. Else, we perform a simple perceptron update on the parameters:

$$\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y') \tag{2}$$

The iteration on the training set is repeated T times, or until convergence. The most expensive step in this algorithm is the calculation of $y' = \text{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$ - this is the *decoding* problem.

Algorithm 1 Online perceptron learning

```
Require: - Set of n training examples $\{(x_i, z_i) : i = 1...n\}$
         - For each incremental step $j$ with $0 \leq j \leq |x_i|$,
           we define the partially parsed utterance $x_i^j$
           and its gold standard semantics $z_i^j$
         - $T$: number of iterations over the training set
         - GEN$(x)$: function enumerating possible parses
           for an input $x$, according to the CCG grammar.
         - GEN$(x, z)$: function enumerating possible parses
           for an input $x$ and which have semantics $z$,
           according to the CCG grammar.
         - $L(y)$ maps a parse tree $y$ to its logical form.
         - Initial parameter vector $\mathbf{w_0}$

  % Initialise
  $\mathbf{w} \leftarrow \mathbf{w_0}$
  % Loop $T$ times on the training examples
  for $t = 1 \ldots T$ do
    for $i = 1 \ldots n$ do
      % Loop on the incremental parsing steps
      for $j = 0...|x_i|$ do
        % Compute best parse according to model
        Let $y' = \text{argmax}_{y \in \mathbf{GEN}(x_i^j)} \mathbf{w}^T \cdot \mathbf{f}(x_i^j, y)$
        % If the decoded parse $\neq$ expected parse, update the
        parameters of the model
        if $L(y') \neq z_i^j$ then
          % Search the best parse for the partial utterance $x_i^j$
          with semantics $z_i^j$
          Let $y^* = \text{argmax}_{y \in \mathbf{GEN}(x_i^j, z_i^j)} \mathbf{w}^T \cdot \mathbf{f}(x_i^j, y)$
          % Update parameter vector $\mathbf{w}$
          Set $\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i^j, y^*) - \mathbf{f}(x_i^j, y')$
        end if
      end for
    end for
  end for
  return parameter vector $\mathbf{w}$
```

It is possible to prove that, provided the training set (x_i, z_i) is separable with margin $\delta > 0$, the algorithm is assured to converge after a finite number of iterations to a model with zero training errors [5]. See also [18] for convergence theorems and proofs.

D. Features

As we have just seen, the parse selection operates by enumerating the possible parses and selecting the one with the highest score according to the linear model parametrised by the weights $\mathbf{w}$.

[2] We used MARY (http://mary.dfki.de) for the text-to-speech engine.

The accuracy of our method crucially relies on the selection of "good" features $\mathbf{f}(x,y)$ for our model - that is, features which help *discriminating* the parses. In our model, the features are of four types: semantic features, syntactic features, contextual features, and speech recognition features.

1) Semantic features: What are the substructures of a logical form which may be relevant to discriminate the parses? We define features on the following information sources: the nominals, the ontological sorts of the nominals, the dependency relations (following [19]), and the sequences of dependency relations.

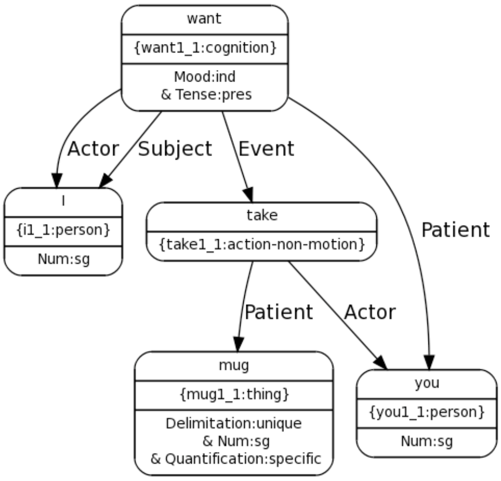


Fig. 3. HLDS logical form for "*I want you to take the mug*".

The features on nominals and ontological sorts aim at modeling (aspects of) *lexical semantics* - e.g. which meanings are the most frequent for a given word -, whereas the features on relations and sequence of relations focus on *sentential semantics* - which dependencies are the most frequent.

These features therefore help us handle various forms of lexical and syntactic ambiguities.

2) Syntactic features: Syntactic features are features associated to the *derivational history* of a specific parse. Alongside the usual CCG rules (application, composition and type raising), our parser also uses a set of non-standard rules designed to handle disfluencies, speech recognition errors, and combinations of discourse units by selectively relaxing the grammatical constraints (see [4] for details). In order to "*penalise*" to a correct extent the application of these non-standard rules, we include in the feature vector $\mathbf{f}(x,y)$ new features counting the number of times these rules are applied in the parse. In the derivation shown in the Figure 4, the rule *corr* (correction of a speech recognition error) is for instance applied once.

These syntactic features can be seen as a *penalty* given to the parses using these non-standard rules, thereby giving a preference to the "normal" parses over them.

This ensures that the grammar relaxation is only applied "as a last resort" when the usual grammatical analysis fails to provide a parse.

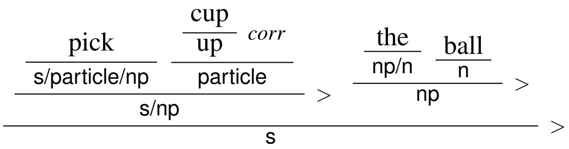


Fig. 4. CCG derivation of *"pick cup the ball"*.

3) Contextual features: As we have already outlined in the background section, one striking characteristic of spoken dialogue is the importance of *context*. Understanding the visual and discourse contexts is crucial to resolve potential ambiguities and compute the most likely interpretation(s) of a given utterance.

The feature vector $\mathbf{f}(x,y)$ therefore includes various features related to the context:

- *Activated words*: our dialogue system maintains in its working memory a list of contextually activated words (cfr. [20]). This list is continuously updated as the dialogue and the environment evolves. For each context-dependent word, we include one feature counting the number of times it appears in the utterance string.
- *Expected dialogue moves*: for each possible dialogue move, we include one feature indicating if the dialogue move is consistent with the current discourse model. These features ensure for instance that the dialogue move following a QuestionYN is a Accept, Reject or another question (e.g. for clarification requests), but almost never an Opening.

4) Speech recognition features: Finally, the feature vector $\mathbf{f}(x,y)$ also includes features related to the *speech recognition*. The ASR module outputs a set of (partial) recognition hypotheses, packed in a word lattice. One example is given in Figure 5.

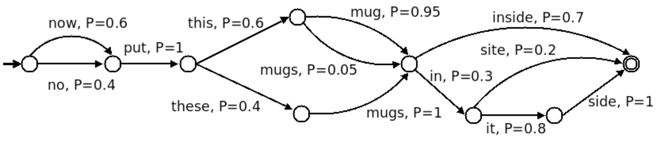


Fig. 5. Example of word lattice

We want to favour the hypotheses with high confidence scores, which are, according to the statistical models incorporated in the ASR, more likely to reflect what was uttered. To this end, we introduce in the feature vector several acoustic features measuring the likelihood of each recognition hypothesis.

E. Incremental chart pruning

In the previous subsections, we explained how the parse selection was performed, and on basis of which features.

	Beam width	Size of word lattice	Average parsing time (in s.)	Exact-match			Partial-match		
				Precision	Recall	F_1-value	Precision	Recall	F_1-value
(Baseline)	*(none)*	*10*	*10.1*	*40.4*	*100.0*	*57.5*	*81.4*	*100.0*	*89.8*
	120	10	5.78	40.9	96.9	57.5	81.9	98.0	89.2
	60	10	**4.82**	41.1	92.5	56.9	81.7	94.1	87.4
	40	10	4.66	39.9	88.1	54.9	79.6	91.9	85.3
	30	10	4.21	41.0	83.0	54.9	80.2	88.6	84.2
	20	10	4.30	40.1	80.3	53.5	78.9	86.5	82.5
(Baseline)	*(none)*	*5*	*5.28*	*40.0*	*100.0*	*57.1*	*81.5*	*100.0*	*89.8*
	120	5	6.62	40.9	98.4	57.8	81.6	98.5	89.3
	60	5	5.28	40.5	96.9	57.1	81.7	97.1	88.7
	40	5	4.26	40.9	91.0	56.5	81.7	92.4	86.7
	30	5	**3.51**	40.7	92.4	56.5	81.4	93.9	87.2
	20	5	2.81	36.7	87.1	51.7	79.6	90.7	84.8

TABLE I

EVALUATION RESULTS (IN SECONDS FOR THE PARSING TIME, IN % FOR THE EXACT- AND PARTIAL-MATCH).

This parse selection is used at each incremental step to discriminate between the ”good” parses that needs to be kept in the parse chart, and the parses that should be pruned in order to keep a limited number of interpretations, and hence avoid a combinatory explosion of analyses.

To achieve this, we introduce a new parameter in our parser: the *beam width*. The beam width defines the maximal number of analyses which can be kept in the chart at each incremental step. If the number of possible readings exceeds the beam width, the analyses with a lower parse selection score are removed from the chart.

Practically, this is realised by removing the top signs associated in the chart with the set of analyses to prune, as well as all the intermediate signs which are included in these top signs *and* are not used in any of the ”good” analyses retained by the parse selection module.

A simple backtracking mechanism is also implemented in the parser. In case the beam width happens to be too narrow and renders the utterance unparsable, it is possible to reintroduce the signs previously removed from the chart and restart the parse at the failure point.

The combination of incremental parsing and incremental chart pruning provides two decisive advantages over classical, non-incremental parsing techniques: first, we can start processing the spoken inputs as soon as a partial analysis can be outputted by the speech recogniser. Second, the pruning mechanism ensures that each incremental parsing step remains time-bounded. Such a combination is therefore ideally suited for the real-time spoken dialogue systems used in human-robot interaction.

IV. EVALUATION

We performed a quantitative evaluation of our approach, using its implementation in a fully integrated system (cf. Section II). To set up the experiments for the evaluation, we have gathered a Wizard-of-Oz corpus of human-robot spoken dialogue for our task-domain (Figure 6), which we segmented and annotated manually with their expected semantic interpretation. The data set contains 195 individual utterances [3] along with their complete logical forms.

The results are shown in the Table I. We tested our approach for five different values of the beam width parameter, and for two sizes of the word lattice. The results are compared against a baseline, which is the performance of our parser without chart pruning. For each configuration, we give the average parsing time, as well as the exact-match and partial-match results (in order to verify that the performance increase is not cancelled by a drop in accuracy). The most important observation we can make is that the choice of the beam width parameter is crucial. Above 30, the chart pruning mechanism works very efficiently – we observe a notable decrease in the parsing time without significantly affecting the accuracy performance. Below 30, the beam width is too small to retain all the necessary information in the chart, and the recall quickly drops.

Figure 7 illustrates the evolution of the ambiguity level (in terms of number of alternative semantic interpretations) during the incremental parsing. We observe that the chart pruning mechanism acts as a *stabilising factor* within the parser, by limiting the number of ambiguities produced after every incremental step to a reasonable level.

Fig. 6. Wizard-of-Oz experiments for a task domain of object manipulation and visual learning

[3] More precisely, word lattices provided by the speech recogniser. These word lattices can contain a maximum of 10 recognition hypotheses.

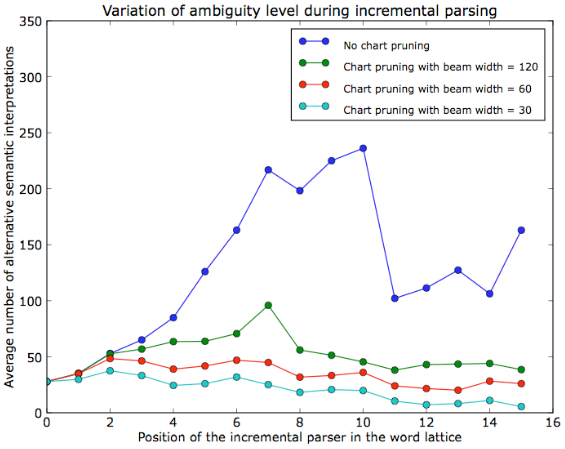


Fig. 7. Variation of ambiguity level during incremental parsing, with and without chart pruning (on word lattices with NBest 10 hypotheses).

V. Conclusions

We presented in this paper an original mechanism for efficient parsing of spoken inputs, based on a combination of incremental *parsing* (to start the processing as soon as a partial speech input is recognised) and incremental *chart pruning* (to limit at every step the number of analyses retained in the parse chart).

The incremental parser is based on a fine-grained Combinatory Categorial Grammar, and takes ASR word lattices as input. It outputs a set of partial semantic interpretations (”logical forms”), which are progressively refined and extended as the utterance unfolds.

Once the partial interpretations are computed, they are subsequently pruned/filtered to keep only the most likely hypotheses in the parse chart. This mechanism is based on a *discriminative model* exploring a set of relevant semantic, syntactic, contextual and acoustic features extracted for each parse. At each incremental step, the discriminative model yields a score for each resulting parse. The parser then only retains in its chart the set of parses associated with a high score, the others being pruned.

The experimental evaluation conducted on a Wizard-of-Oz test suite demonstrated that the aforementioned approach was able to significantly improve the parser performance .

As forthcoming work, we shall examine the extension of our approach in new directions, such as the introduction of more refined contextual features, the extension of the grammar relaxation rules, or the use of more sophisticated learning algorithms such as Support Vector Machines.

References

[1] R. Fernández and J. Ginzburg, “A corpus study of non-sentential utterances in dialogue,” *Traitement Automatique des Langues*, vol. 43, no. 2, pp. 12–43, 2002.

[2] E. A. Topp, H. Hüttenrauch, H. Christensen, and K. Severinson Eklundh, “Bringing together human and robotic environment representations – a pilot study,” in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October 2006.

[3] P. Lison, “Robust processing of situated spoken dialogue,” Master’s thesis, Universität des Saarlandes, Saarbrücken, 2008, http://www.dfki.de/∼ plison/pubs/thesis/main.thesis.plison2008.pdf.

[4] P. Lison and G.-J. M. Kruijff, “An integrated approach to robust processing of situated spoken dialogue,” in *Proceedings of the International Workshop on Semantic Representation of Spoken Language (SRSL’09)*, Athens, Greece, 2009, (to appear).

[5] M. Collins and B. Roark, “Incremental parsing with the perceptron algorithm,” in *ACL ’04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 111.

[6] N. A. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. M. Kruijff, M. Brenner, G. Berginc, and D. Skocaj, “Towards an integrated robot with multiple cognitive functions.” in *Proc. AAAI’07*. AAAI Press, 2007, pp. 1548–1553.

[7] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes, “Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction,” in *Language and Robots: Proceedings from the Symposium (LangRo’2007)*, Aveiro, Portugal, December 2007, pp. 55–64.

[8] M. Steedman and J. Baldridge, “Combinatory categorial grammar,” in *Nontransformational Syntax: A Guide to Current Models*, R. Borsley and K. Börjars, Eds. Oxford: Blackwell, 2009.

[9] J. Baldridge and G.-J. M. Kruijff, “Coupling CCG and hybrid logic dependency semantics,” in *ACL’02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics, 2002, pp. 319–326.

[10] T. Kasami, “An efficient recognition and syntax analysis algorithm for context free languages,” Air Force Cambridge Research Laboratory, Bedford, Massachussetts, Scientific Report AF CRL-65-758, 1965.

[11] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, “Crossmodal content binding in information-processing architectures,” in *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008.

[12] P. Knoeferle and M. Crocker, “The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking,” *Cognitive Science*, 2006.

[13] D. Roy, “Semiotic schemas: A framework for grounding language in action and perception,” *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[14] D. Roy and N. Mukherjee, “Towards situated speech understanding: visual context priming of language models,” *Computer Speech & Language*, vol. 19, no. 2, pp. 227–248, April 2005.

[15] T. Brick and M. Scheutz, “Incremental natural language processing for HRI,” in *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI’07)*, 2007, pp. 263 – 270.

[16] K. Weilhammer, M. N. Stuttle, and S. Young, “Bootstrapping language models for dialogue systems,” in *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA, 2006.

[17] L. S. Zettlemoyer and M. Collins, “Online learning of relaxed CCG grammars for parsing to logical form,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 678–687.

[18] M. Collins, “Parameter estimation for statistical parsing models: theory and practice of distribution-free methods,” in *New developments in parsing technology*. Kluwer Academic Publishers, 2004, pp. 19–55.

[19] S. Clark and J. R. Curran, “Log-linear models for wide-coverage ccg parsing,” in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 97–104.

[20] P. Lison and G.-J. M. Kruijff, “Salience-driven contextual priming of speech recognition for human-robot interaction,” in *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras (Greece), 2008.

An Integrated Approach to Robust Processing of Situated Spoken Dialogue

Pierre Lison
Language Technology Lab,
DFKI GmbH,
Saarbrücken, Germany
pierre.lison@dfki.de

Geert-Jan M. Kruijff
Language Technology Lab,
DFKI GmbH,
Saarbrücken, Germany
gj@dfki.de

Abstract

Spoken dialogue is notoriously hard to process with standard NLP technologies. Natural spoken dialogue is replete with disfluent, partial, elided or ungrammatical utterances, all of which are difficult to accommodate in a dialogue system. Furthermore, speech recognition is known to be a highly error-prone task, especially for complex, open-ended domains. The combination of these two problems – ill-formed and/or misrecognised speech inputs – raises a major challenge to the development of robust dialogue systems.

We present an integrated approach for addressing these two issues, based on an incremental parser for Combinatory Categorial Grammar. The parser takes word lattices as input and is able to handle ill-formed and misrecognised utterances by selectively relaxing its set of grammatical rules. The choice of the most relevant interpretation is then realised via a discriminative model augmented with contextual information. The approach is fully implemented in a dialogue system for autonomous robots. Evaluation results on a Wizard of Oz test suite demonstrate very significant improvements in accuracy and robustness compared to the baseline.

1 Introduction

Spoken dialogue is often considered to be one of the most natural means of interaction between a human and a robot. It is, however, notoriously hard to process with standard language processing technologies. Dialogue utterances are often incomplete or ungrammatical, and may contain numerous disfluencies like fillers (err, uh, mm), repetitions, self-corrections, etc. Rather than getting crisp-and-clear commands such as ”*Put the red ball inside the box!*”, it is more likely the robot will hear such kind of utterance: ”*right, now, could you, uh, put the red ball, yeah, inside the ba/ box!*”. This is natural behaviour in human-human interaction (Fernández and Ginzburg, 2002) and can also be observed in several domain-specific corpora for human-robot interaction (Topp et al., 2006).

Moreover, even in the (rare) case where the utterance is perfectly well-formed and does not contain any kind of disfluencies, the dialogue system still needs to accomodate the various speech recognition errors thay may arise. This problem is particularly acute for robots operating in real-world noisy environments and deal with utterances pertaining to complex, open-ended domains.

The paper presents a new approach to address these two difficult issues. Our starting point is the work done by Zettlemoyer and Collins on parsing using relaxed CCG grammars (Zettlemoyer and Collins, 2007) (ZC07). In order to account for natural spoken language phenomena (more flexible word order, missing words, etc.), they augment their grammar framework with a small set of non-standard combinatory rules, leading to a *relaxation* of the grammatical constraints. A discriminative model over the parses is coupled with the parser, and is responsible for selecting the most likely interpretation(s) among the possible ones.

In this paper, we extend their approach in two important ways. First, ZC07 focused on the treatment of ill-formed input, and ignored the speech recognition issues. Our system, to the contrary, is able to deal with both ill-formed and misrecognized input, in an integrated fashion. This is done by augmenting the set of non-standard combinators with new rules specifically tailored to deal with speech recognition errors.

Second, the only features used by ZC07 are syntactic features (see section 3.4 for details). We significantly extend the range of features included in the discriminative model, by incorporating not only *syntactic*, but also *acoustic*, *semantic* and *contextual* information into the model. As the experimental results have shown, the inclusion of a broader range of linguistic and contextual information leads to a more accurate discrimination of the various interpretations.

An overview of the paper is as follows. We first describe in Section 2 the cognitive architecture in which our system has been integrated. We then discuss the approach in detail in Section 3. Finally, we present in Section 4 the quantitative evaluations on a WOZ test suite, and conclude.

2 Architecture

The approach we present in this paper is fully implemented and integrated into a cognitive architecture for autonomous robots. A recent version of this system is described in (Hawes et al., 2007). It is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks.

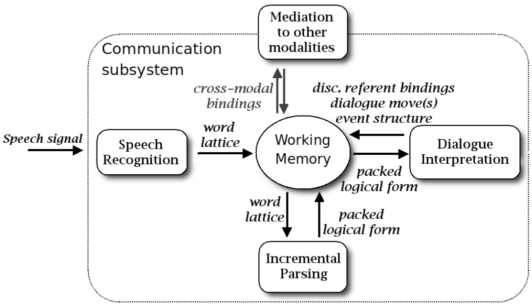


Figure 1: Architecture schema of the communication subsystem (only for comprehension).

Figure 2 illustrates the architecture schema for the communication subsystem incorporated in the cognitive architecture (only the comprehension part is shown).

Starting with ASR, we process the audio signal to establish a *word lattice* containing statistically ranked hypotheses about word sequences. Subsequently, parsing constructs grammatical analyses for the given word lattice. A grammatical analysis constructs both a syntactic analysis of the utterance, and a representation of its meaning. The analysis is based on an incremental chart parser[1] for Combinatory Categorial Grammar (Steedman and Baldridge, 2009). These meaning representations are ontologically richly sorted, relational structures, formulated in a (propositional) description logic, more precisely in the HLDS formalism (Baldridge and Kruijff, 2002). The parser compacts all meaning representations into a single *packed logical form* (Carroll and Oepen, 2005; Kruijff et al., 2007). A packed LF represents content similar across the different analyses as a single graph, using over- and underspecification of how different nodes can be connected to capture lexical and syntactic forms of ambiguity.

At the level of dialogue interpretation, a packed logical form is resolved against a SDRS-like dialogue model (Asher and Lascarides, 2003) to establish co-reference and dialogue moves.

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* accross modalities (Jacobsson et al., 2008).

2.1 Context-sensitivity

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of utterance as it unfolds (Knoeferle and Crocker, 2006). During utterance comprehension, humans combine linguistic information with scene understanding and "world knowledge".

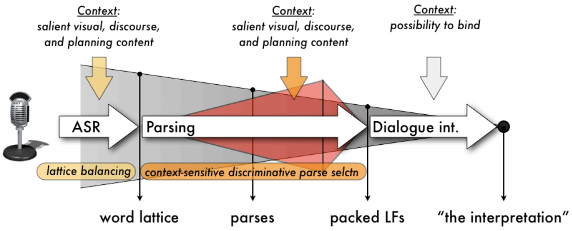

Context:
salient visual, discourse, and planning content
Context:
salient visual, discourse, and planning content
Context:
possibility to bind
ASR
Parsing
Dialogue int.
lattice balancing
context-sensitive discriminative parse selctn
word lattice
parses
packed LFs
"the interpretation"

Figure 2: Context-sensitivity in processing situated dialogue understanding

[1] Built using the OpenCCG API: http://openccg.sf.net

Several approaches in situated dialogue for human-robot interaction have made similar observations (Roy, 2005; Roy and Mukherjee, 2005; Brick and Scheutz, 2007; Kruijff et al., 2007): A robot's understanding can be improved by relating utterances to the situated context. As we will see in the next section, by incorporating contextual information into our model, our approach to robust processing of spoken dialogue seeks to exploit this important insight.

3 Approach

3.1 Grammar relaxation

Our approach to robust processing of spoken dialogue rests on the idea of **grammar relaxation**: the grammatical constraints specified in the grammar are "relaxed" to handle slightly ill-formed or misrecognised utterances.

Practically, the grammar relaxation is done via the introduction of *non-standard CCG rules* (Zettlemoyer and Collins, 2007). In Combinatory Categorial Grammar, the rules are used to assemble categories to form larger pieces of syntactic and semantic structure. The standard rules are application ($<, >$), composition ($\mathbf{B}$), and type raising ($\mathbf{T}$) (Steedman and Baldridge, 2009).

Several types of non-standard rules have been introduced. We describe here the two most important ones: the *discourse-level composition rules*, and the *ASR correction rules*. We invite the reader to consult (Lison, 2008) for more details on the complete set of grammar relaxation rules.

3.1.1 Discourse-level composition rules

In natural spoken dialogue, we may encounter utterances containing several independent "chunks" without any explicit separation (or only a short pause or a slight change in intonation), such as

(1) "yes take the ball no the other one on your left right and now put it in the box."

Even if retrieving a fully structured parse for this utterance is difficult to achieve, it would be useful to have access to a list of smaller "discourse units". Syntactically speaking, a discourse unit can be any type of saturated atomic categories - from a simple discourse marker to a full sentence.

The type-changing rule $\mathbf{T}_{du}$ allows the conversion of atomic categories into discourse units:

$$\mathsf{A} : @_i f \Rightarrow \mathsf{du} : @_i f \qquad (\mathbf{T}_{du})$$

where A represents an arbitrary saturated atomic category (s, np, pp, etc.).

The rule $\mathbf{T}_C$ is a type-changing rule which allows us to integrate two discourse units into a single structure:

$$\mathsf{du} : @_a x \Rightarrow \mathsf{du} : @_c z \;/\; \mathsf{du} : @_b y \qquad (\mathbf{T}_C)$$

where the formula $@_c z$ is defined as:

$$@_{\{c:\text{d-units}\}}(\mathbf{list} \wedge (\langle \text{FIRST} \rangle\ a \wedge x) \wedge (\langle \text{NEXT} \rangle\ b \wedge y)) \qquad (2)$$

3.1.2 ASR error correction rules

Speech recognition is a highly error-prone task. It is however possible to partially alleviate this problem by inserting new error-correction rules (more precisely, new lexical entries) for the most frequently misrecognised words.

If we notice e.g. that the ASR system frequently substitutes the word "wrong" for the word "round" during the recognition (because of their phonological proximity), we can introduce a new lexical entry in the lexicon in order to correct this error:

$$\textit{round} \vdash \mathsf{adj} : @_{\textit{attitude}}(\mathbf{wrong}) \qquad (3)$$

A set of thirteen new lexical entries of this type have been added to our lexicon to account for the most frequent recognition errors.

3.2 Parse selection

Using more powerful grammar rules to relax the grammatical analysis tends to increase the number of parses. We hence need a mechanism to discriminate among the possible parses. The task of selecting the most likely interpretation among a set of possible ones is called *parse selection*. Once all the possible parses for a given utterance are computed, they are subsequently filtered or selected in order to retain only the most likely interpretation(s). This is done via a (discriminative) statistical model covering a large number of features.

Formally, 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. We assume:

1. A function $\mathbf{GEN}(x)$ which enumerates all possible parses for an input x. In our case, this function simply represents the set of 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 $\mathbf{w} \in \Re^d$.

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

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \ \mathbf{w}^T \cdot \mathbf{f}(x, y) \qquad (4)$$

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

The task of parse selection is an example of a *structured classification problem*, which is the problem of predicting an output y from an input x, where the output y has a rich internal structure. In the specific case of parse selection, x is a word lattice, and y a logical form.

3.3 Learning

3.3.1 Training data

In order to estimate the parameters $\mathbf{w}$, we need a set of training examples. Unfortunately, no corpus of situated dialogue adapted to our task domain is available to this day, let alone semantically annotated. The collection of in-domain data via Wizard of Oz experiments being a very costly and time-consuming process, we followed the approach advocated in (Weilhammer et al., 2006) and *generated* a corpus from a hand-written task grammar.

To this end, we first collected a small set of WoZ data, totalling about a thousand utterances. This set is too small to be directly used as a corpus for statistical training, but sufficient to capture the most frequent linguistic constructions in this particular context. Based on it, we designed a domain-specific CFG grammar covering most of the utterances. Each rule is associated to a semantic HLDS representation. Weights are automatically assigned to each grammar rule by parsing our corpus, hence leading to a small *stochastic CFG grammar* augmented with semantic information.

Once the grammar is specified, it is randomly traversed a large number of times, resulting in a larger set (about 25.000) of utterances along with their semantic representations. Since we are interested in handling errors arising from speech recognition, we also need to "simulate" the most frequent recognition errors. To this end, we *synthesise* each string generated by the domain-specific CFG grammar, using a text-to-speech engine[2], feed the audio stream to the speech recogniser, and retrieve the recognition result. Via this technique, we are able to easily collect a large amount of training data[3].

3.3.2 Perceptron learning

The algorithm we use to estimate the parameters $\mathbf{w}$ using the training data is a **perceptron**. The algorithm is fully online - it visits each example in turn and updates $\mathbf{w}$ if necessary. Albeit simple, the algorithm has proven to be very efficient and accurate for the task of parse selection (Collins and Roark, 2004; Collins, 2004; Zettlemoyer and Collins, 2005; Zettlemoyer and Collins, 2007).

The pseudo-code for the online learning algorithm is detailed in [**Algorithm 1**].

It works as follows: the parameters $\mathbf{w}$ are first initialised to some arbitrary values. Then, for each pair (x_i, z_i) in the training set, the algorithm searchs for the parse y' with the highest score according to the current model. If this parse happens to match the best parse which generates z_i (which we shall denote y^*), we move to the next example. Else, we perform a simple perceptron update on the parameters:

$$\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y') \qquad (5)$$

The iteration on the training set is repeated T times, or until convergence.

The most expensive step in this algorithm is the calculation of $y' = \operatorname{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$ - this is the *decoding* problem.

It is possible to prove that, provided the training set (x_i, z_i) is separable with margin $\delta > 0$, the

[2] We used MARY (http://mary.dfki.de) for the text-to-speech engine.

[3] Because of its relatively artificial character, the quality of such training data is naturally lower than what could be obtained with a genuine corpus. But, as the experimental results will show, it remains sufficient to train the perceptron for the parse selection task, and achieve significant improvements in accuracy and robustness. In a near future, we plan to progressively replace this generated training data by a real spoken dialogue corpus adapted to our task domain.

algorithm is assured to converge after a finite number of iterations to a model with zero training errors (Collins and Roark, 2004). See also (Collins, 2004) for convergence theorems and proofs.

Algorithm 1 Online perceptron learning

```
Require: - set of $n$ training examples $\{(x_i, z_i) : i = 1...n\}$
         - $T$: number of iterations over the training set
         - $\texttt{GEN}(x)$: function enumerating possible parses
           for an input $x$, according to the CCG grammar.
         - $\texttt{GEN}(x, z)$: function enumerating possible parses
           for an input $x$ and which have semantics $z$,
           according to the CCG grammar.
         - $L(y)$ maps a parse tree $y$ to its logical form.
         - Initial parameter vector $\mathbf{w_0}$

  % Initialise
  $\mathbf{w} \leftarrow \mathbf{w_0}$
  % Loop $T$ times on the training examples
  for $t = 1...T$ do
    for $i = 1...n$ do
      % Compute best parse according to current model
      Let $y' = \text{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$
      % If the decoded parse $\neq$ expected parse, update the parameters
      if $L(y') \neq z_i$ then
        % Search the best parse for utterance $x_i$ with semantics $z_i$
        Let $y^* = \text{argmax}_{y \in \mathbf{GEN}(x_i, z_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$
        % Update parameter vector $\mathbf{w}$
        Set $\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y')$
      end if
    end for
  end for
  return parameter vector $\mathbf{w}$
```

3.4 Features

As we have seen, the parse selection operates by enumerating the possible parses and selecting the one with the highest score according to the linear model parametrised by $\mathbf{w}$.

The accuracy of our method 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: semantic features, syntactic features, contextual features, and speech recognition features.

3.4.1 Semantic features

What are the substructures of a logical form which may be relevant to discriminate the parses? We define features on the following information sources:

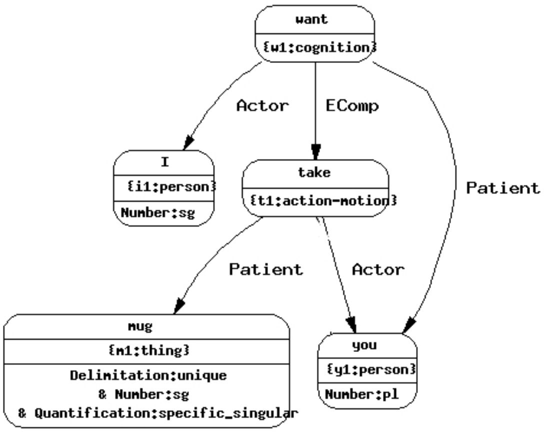


Figure 3: graphical representation of the HLDS logical form for "*I want you to take the mug*".

1. *Nominals*: for each possible pair $\langle prop, sort \rangle$, we include a feature f_i in $\mathbf{f}(x, y)$ counting the number of nominals with ontological sort $sort$ and proposition $prop$ in the logical form.

2. *Ontological sorts*: occurrences of specific ontological sorts in the logical form.

3. *Dependency relations*: following (Clark and Curran, 2003), we also model the *dependency structure* of the logical form. Each dependency relation is defined as a triple $\langle sort_a,\ sort_b,\ label \rangle$, where $sort_a$ denotes the sort of the incoming nominal, $sort_b$ the sort of the outgoing nominal, and $label$ is the relation label.

4. *Sequences of dependency relations*: number of occurrences of particular sequences (ie. bigram counts) of dependency relations.

The features on nominals and ontological sorts aim at modeling (aspects of) *lexical semantics* - e.g. which meanings are the most frequent for a given word -, whereas the features on relations and sequence of relations focus on *sentential semantics* - which dependencies are the most frequent. These features therefore help us handle lexical and syntactic ambiguities.

3.4.2 Syntactic features

By "syntactic features", we mean features associated to the *derivational history* of a specific parse. The main use of these features is to *penalise* to a correct extent the application of the non-standard rules introduced into the grammar.

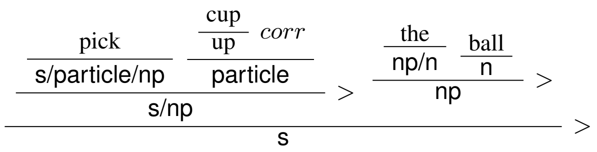


Figure 4: CCG derivation of *"pick cup the ball"*.

To this end, we include in the feature vector $\mathbf{f}(x,y)$ a new feature for each non-standard rule, which counts the number of times the rule was applied in the parse.

In the derivation shown in the figure 4, the rule *corr* (correction of a speech recognition error) is applied once, so the corresponding feature value is set to 1. The feature values for the remaining rules are set to 0, since they are absent from the parse.

These syntactic features can be seen as a *penalty* given to the parses using these non-standard rules, thereby giving a preference to the "normal" parses over them. This mechanism ensures that the grammar relaxation is only applied "as a last resort" when the usual grammatical analysis fails to provide a full parse. Of course, depending on the relative frequency of occurrence of these rules in the training corpus, some of them will be more strongly penalised than others.

3.4.3 Contextual features

As we have already outlined in the background section, one striking characteristic of spoken dialogue is the importance of *context*. Understanding the visual and discourse contexts is crucial to resolve potential ambiguities and compute the most likely interpretation(s) of a given utterance.

The feature vector $\mathbf{f}(x,y)$ therefore includes various features related to the context:

1. *Activated words*: our dialogue system maintains in its working memory a list of contextually activated words (cfr. (Lison and Kruijff, 2008)). This list is continuously updated as the dialogue and the environment evolves. For each context-dependent word, we include one feature counting the number of times it appears in the utterance string.

2. *Expected dialogue moves*: for each possible dialogue move, we include one feature indicating if the dialogue move is consistent with the current discourse model. These features ensure for instance that the dialogue move following a QuestionYN is a Accept, Reject or another question (e.g. for clarification requests), but almost never an Opening.

3. *Expected syntactic categories*: for each atomic syntactic category in the CCG grammar, we include one feature indicating if the category is consistent with the current discourse model. These features can be used to handle *sentence fragments*.

3.4.4 Speech recognition features

Finally, the feature vector $\mathbf{f}(x,y)$ also includes features related to the *speech recognition*. The ASR module outputs a set of (partial) recognition hypotheses, packed in a word lattice. One example of such a structure is given in Figure 5. Each recognition hypothesis is provided with an associated confidence score, and we want to favour the hypotheses with high confidence scores, which are, according to the statistical models incorporated in the ASR, more likely to reflect what was uttered.

To this end, we introduce three features: the *acoustic confidence score* (confidence score provided by the statistical models included in the ASR), the *semantic confidence score* (based on a "concept model" also provided by the ASR), and the *ASR ranking* (hypothesis rank in the word lattice, from best to worst).

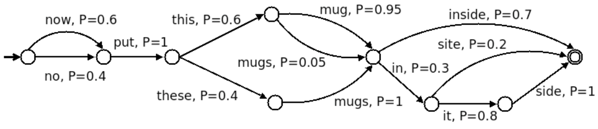


Figure 5: Example of word lattice

4 Experimental evaluation

We performed a quantitative evaluation of our approach, using its implementation in a fully integrated system (cf. Section 2). To set up the experiments for the evaluation, we have gathered a corpus of human-robot spoken dialogue for our task-domain, which we segmented and annotated manually with their expected semantic interpretation. The data set contains 195 individual utterances along with their complete logical forms.

4.1 Results

Three types of quantitative results are extracted from the evaluation results: *exact-match*, *partial-*

	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 1: 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 2: Partial-match accuracy results (in percents).

match, and *word error rate*. Tables 1, 2 and 3 illustrate the results, broken down by use of grammar relaxation, use of parse selection, and number of recognition hypotheses considered.

Each line in the tables corresponds to a possible configuration. Tables 1 and 2 give the precision, recall and F_1 value for each configuration (respectively for the exact- and partial-match), and Table 3 gives the Word Error Rate [WER].

The first line corresponds to the baseline: no grammar relaxation, no parse selection, and use of the first NBest recognition hypothesis. The last line corresponds to the results with the full approach: grammar relaxation, parse selection, and use of 10 recognition hypotheses.

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 3: Word error rate (in percents).

4.2 Comparison with baseline

Here are the comparative results we obtained:

- Regarding the exact-match results between the baseline and our approach (grammar relaxation and parse selection with all features activated for NBest 10), the F_1-measure climbs from 43.0 % to 67.2 %, which means a relative difference of $\mathbf{56.3}$ %.
- For the partial-match, the F_1-measure goes from 68.0 % for the baseline to 87.3 % for our approach – a relative increase of $\mathbf{28.4}$ %.
- We observe a significant decrease in WER: we go from 20.5 % for the baseline to 15.7 % with our approach. The difference is statistically significant (p-value for t-tests is 0.036), and the relative decrease of $\mathbf{23.4}$ %.

5 Conclusions

We presented an *integrated* approach to the processing of (situated) spoken dialogue, suited to the specific needs and challenges encountered in human-robot interaction.

In order to handle disfluent, partial, ill-formed or misrecognized utterances, the grammar used by the parser is "relaxed" via the introduction of a set of *non-standard combinators* which allow for the insertion/deletion of specific words, the combination of discourse fragments or the correction of speech recognition errors.

The relaxed parser yields a (potentially large) set of parses, which are then packed and retrieved by the parse selection module. The parse selection is based on a discriminative model exploring a set of relevant semantic, syntactic, contextual and acoustic features extracted for each parse. The parameters of this model are estimated against an automatically generated corpus of $\langle$utterance, logical form$\rangle$ pairs. The learning algorithm is an perceptron, a simple albeit efficient technique for parameter estimation.

As forthcoming work, we shall examine the potential extension of our approach in new directions, such as the exploitation of parse selection for *incremental* scoring/pruning of the parse chart, the introduction of more refined contextual features, or the use of more sophisticated learning algorithms, such as Support Vector Machines.

6 Acknowledgements

This work was supported by the EU FP7 ICT Integrated Project "CogX" (FP7-ICT- 215181).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

J. Baldridge and G.-J. M. Kruijff. 2002. Coupling CCG and hybrid logic dependency semantics. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Philadelphia, PA. Association for Computational Linguistics.

T. Brick and M. Scheutz. 2007. Incremental natural language processing for HRI. In *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07)*, pages 263 – 270.

J. Carroll and S. Oepen. 2005. High efficiency realization for a wide-coverage unification grammar. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 165–176.

S. Clark and J. R. Curran. 2003. Log-linear models for wide-coverage ccg parsing. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 97–104, Morristown, NJ, USA. Association for Computational Linguistics.

M. Collins and B. Roark. 2004. Incremental parsing with the perceptron algorithm. In *ACL '04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 111, Morristown, NJ, USA. Association for Computational Linguistics.

M. Collins. 2004. Parameter estimation for statistical parsing models: theory and practice of distribution-free methods. In *New developments in parsing technology*, pages 19–55. Kluwer Academic Publishers.

R. Fernández and J. Ginzburg. 2002. A corpus study of non-sentential utterances in dialogue. *Traitement Automatique des Langues*, 43(2):12–43.

N. A. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. M. Kruijff, M. Brenner, G. Berginc, and D. Skocaj. 2007. Towards an integrated robot with multiple cognitive functions. In *Proc. AAAI'07*, pages 1548–1553. AAAI Press.

H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt. 2008. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15.

P. Knoeferle and M.C. Crocker. 2006. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*.

G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N.A. Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, pages 55–64, Aveiro, Portugal, December.

P. Lison and G.-J. M. Kruijff. 2008. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras (Greece).

P. Lison. 2008. Robust processing of situated spoken dialogue. Master's thesis, Universität des Saarlandes, Saarbrücken. http://www.dfki.de/ plison/pubs/thesis/main.thesis.plison2008.pdf.

D. Roy and N. Mukherjee. 2005. Towards situated speech understanding: visual context priming of language models. *Computer Speech & Language*, 19(2):227–248, April.

D. Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

M. Steedman and J. Baldridge. 2009. Combinatory categorial grammar. In Robert Borsley and Kersti Börjars, editors, *Nontransformational Syntax: A Guide to Current Models*. Blackwell, Oxford.

E. A. Topp, H. Hüttenrauch, H.I. Christensen, and K. Severinson Eklundh. 2006. Bringing together human and robotic environment representations – a pilot study. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October.

K. Weilhammer, M. N. Stuttle, and S. Young. 2006. Bootstrapping language models for dialogue systems. In *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA.

L. S. Zettlemoyer and M. Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 658–666.

L. S. Zettlemoyer and M. Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687.