Modelling context awareness for a situated semantic agent

Vossen, Piek; Bajčetić, Lenka; Baez, Selene; Bašić, Suzana; Kraaijeveld, Bram

published in
Modeling and Using Context
2019

DOI (link to publisher)
10.1007/978-3-030-34974-5_20

document version
Publisher's PDF, also known as Version of record

Modelling Context Awareness for a Situated Semantic Agent

Piek Vossen(✉), Lenka Bajčetić, Selene Baez, Suzana Bašić, and Bram Kraaijeveld

Vrije Universiteit, Amsterdam, The Netherlands
piek.vossen@vu.nl

Abstract. This paper presents a model of contextual awareness implemented for a social communicative robot *Leolani*. Our model starts from the assumption that robots and humans need to establish a common ground about the world they share. This is not trivial as robots make many errors and start with little knowledge. As such, the context in which communication takes place can both help and complicate the interaction: if the context is interpreted correctly it helps in disambiguating the signals, but if it is interpreted wrongly it may distort interpretation. We defined the surrounding world as a spatial context, the communication as a discourse context and the interaction as a social context, which are all three interconnected and have an impact on each other. We model the result of the interpretations as symbolic knowledge (RDF) in a triple store to reason over the result, detect conflicts, uncertainty and gaps. We explain how our model tries to combine the contexts and the signal interpretation and we mention future directions of research to improve this complex process.

Keywords: Robotics · Situated context · Social context · Discourse context

1 Introduction

Without context, we are lost in semantic space. Ambiguity and variation of natural language is so big that meaning is unsolvable without context. Contexts can be defined as knowledge-rich data points that can help in interpreting signals, while signals are structures that convey information. Contexts have predictive power, as they can predict the signal before it is present or when it is masked. Strong evidence for this predictive power comes from current word embedding models trained from Big Data such as Word2Vec [15] or GloVe [16] which predict the direct linguistic context. Embedding models can also be reversed to predict the linguistic context from the signal, showing that the relation is mutual since contexts need to be interpreted as well: contexts define signals and signals define contexts. The difference is more a matter of relevance and focus.

In real-world situations the context can be complex or confusing when interpreted wrongly. This is apparent when modeling (mutual) understanding in

G. Bella and P. Bouquet (Eds.): CONTEXT 2019, LNAI 11939, pp. 238–252, 2019.
https://doi.org/10.1007/978-3-030-34974-5_20

human-robot communication. Robots have limited capacity to perceive and interpret the context in its full complexity. This severely complicates communication. Our robot model therefore aims to take uncertainty and alternative interpretations as a starting point, while using communication to adapt, correct, confirm and reach consensus about interpretation. Possibly wrong interpretations of signals are stored in a 'brain' (an RDF[1] triple store) as symbolic knowledge representations using a Theory-of-Mind model [12] that keeps track of the sources of interpretation and its status. This architecture allows our robot to reason over contextualized signal interpretations and to proactively resolve errors, conflicts, uncertainties and gaps using natural language dialogue.

In this paper, we report on our vision to model contexts of human language communication in real-world situations by building a robot model that communicates about the world and about us. In our previous work [18,19] we introduced the social robot *Leolani* or *L* as a multi-modal semantic agent that uses communication to learn. In this paper, we explain her contextual awareness along three different data layers: the discourse, the surrounding space and the social relationship. We describe how the interpretation of signals and contexts influence each other. In Sect. 2, we explain the foundations for our robot model, while in Sect. 3 we describe the overall robot model. In Sects. 4.1, 4.2 and 4.3, we describe the discourse, spatial and social contexts, respectively, and the way in which our model deals with ambiguity and conflicts in interpreting signals and acquired knowledge. We conclude and look forward in Sect. 6.

2 Background

Our research focuses on *identity*, *reference* and *perspective*. Identity conveys the ability to determine the world we share and the objects, situations and properties within it. Robots perceive the world differently from us and have difficulty identifying situations, entities and properties of the physical world. Identity of the world can therefore be very different across humans and robots. Language commonly makes reference to the physical world and the entities within it. We can make reference to the same things in different ways and different things can be referred to in the same way [6]. We can observe large ambiguity and variation in making reference to the world. While ambiguity can be resolved by context, variation can only be explained by both the context and the perspective of the source that makes reference (a personal context). Perspective can be defined as the personal and social position of the discourse participant with respect to the topic of communication. This position can be defined spatially, where the discourse participants stand with respect to perceived objects, and also socially: what knowledge is shared, what emotions and intentions you have, what the relationship is between the discourse participants.

Identity, reference and perspective are clearly related and ambiguity and variation can only be resolved and explained when dealing with all three aspects in combination. This relation is contextual in nature. What we distinguish and

[1] Resource Description Framework.

consider relevant in the real world is partly determined by our perspective, e.g. intentions and past experiences. The distinguished and relevant world determines references and resolves ambiguity. Such contexts are also defined and established during conversations, which creates a referential context that can be exploited.

In the next section, we will investigate the context created in discourse, the spatial context as the result of awareness, and the way the social context can be established and exploited during conversation between robots and people. We follow a pragmatic modeling approach to study the relation between contexts and the above phenomena, where we also investigate the capacity of the robot to collaborate in reaching an optimal solution. Our robot implementation thus demonstrates context in its full complexity and shows directions of future research to explore. In the next Sect. 3, we first explain the basics of our robot model L .

3 The Robot Model

L is a curious robot equipped with cognitive abilities and communication skills to support social behaviour. When switched on, L scans the objects and people in her environment and relates them to a new instantiated context. Next, she tries to determine her location either by reasoning over previous contexts or asking an available source. Upon encountering people, L tries to discern whether the person is already known and should be greeted as such, or the person is encountered for the first time, in which case a get-to-know sequence is initialised. Subsequently, the robot waits for the person to initiate conversation by asking a question or making a statement. Questions trigger simple (SPARQL) queries, while statements are processed to represent new knowledge along with its provenance. When new information is added, this generates *thoughts*, which are reflections of the current state of the "brain" (the storage of her knowledge) and how this is affected by the newly added information. Through these thoughts the robot raises pro-active questions or comments to the person to improve the state-of-the-brain. These initiatives to improve the state-of-the-brain are defined as inner *drives*, some of which try to harmonize knowledge in relation to the context.

The overall robot model architecture is shown in Fig. 1. We defined four layers: (1) sensor processing layer, (2) communication layer that responds to sensor input or inner drives, (3) language processing layer which deals with questions and statements, and (4) knowledge layer that queries or stores the result of communication, or accesses the Web. We utilize several ready-made modules in the sensor processing layer: WebRTC [2] for speech detection, the Inception neural network [17] for object recognition, OpenFace [1] for face recognition, and Google Cloud Speech-to-Text API [7] for speech recognition. The outputs of these processing modules are used as inputs to the other layers. Hence, we do not address potential conflicts and ambiguities in the signal layer itself, but try to resolve them in the higher-level layers.

Signals are processed either as perceptions of the surroundings or as communication. Visual perceptions are interpreted by object recognition and face

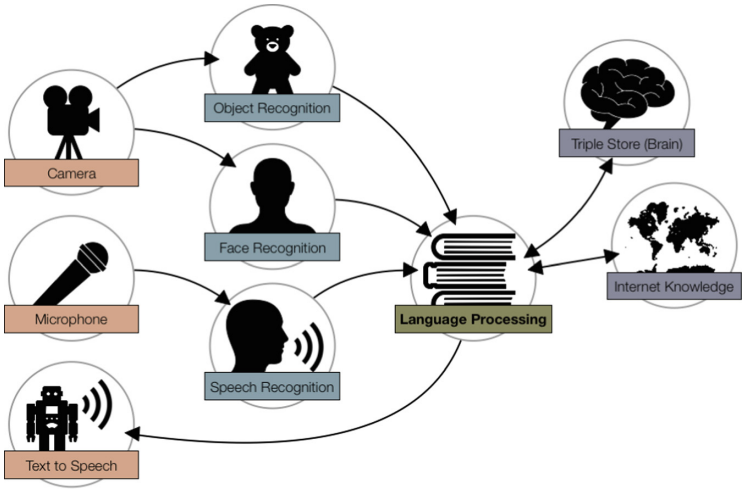

Fig. 1. Overall architecture of the robot model

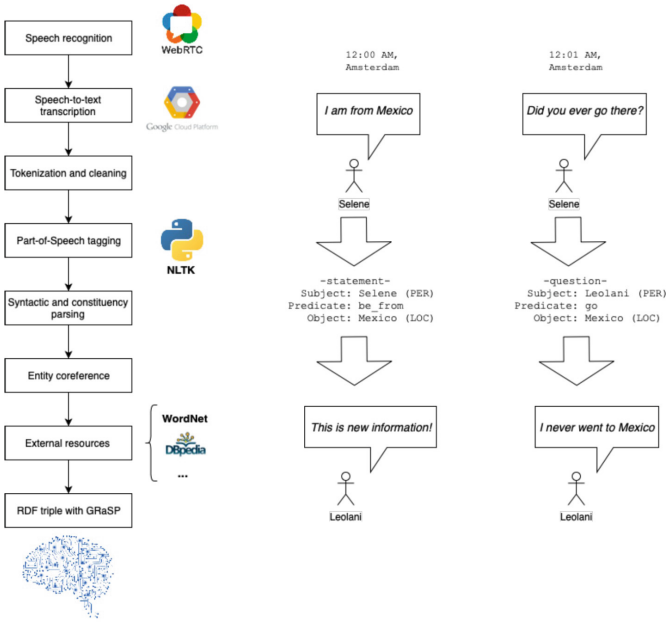

Fig. 2. Natural language processing pipeline

recognition. These are used to define the surrounding context and to identify the people in it. Audio perceptions are processed as language. The result of interaction is stored in an RDF triple store ("the brain"), which stores all interpretations of experiences. The brain forms the basis for the drives of the robot to communicate.

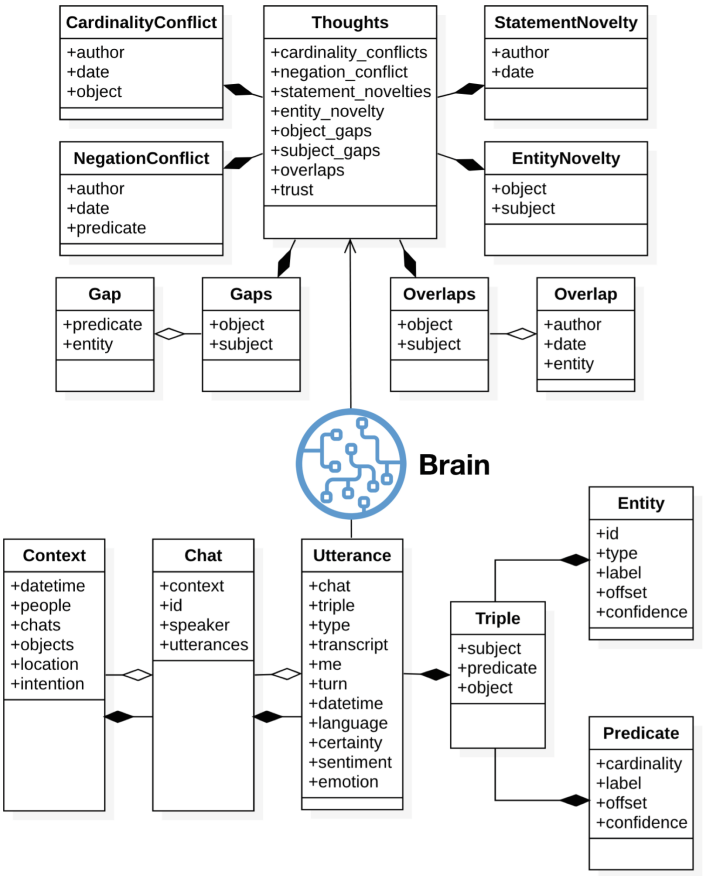

Fig. 3. Data model class diagram

Figure 2, shows the NLP pipeline that is used to process the audio signal. As shown in Fig. 2, the NLP Pipeline consists of several external components, while some are manually implemented specifically for this task. For the sake of transparency, reasoning and control, we resorted to rule-based parsing instead of a neural-net approach. This refers specifically to the syntactic and constituency parsing.

We use the Grounded Representation and Source Perspective (GRaSP) model [5] as a basis for representing content, communication and sources. We have adapted GRaSP to deal with perception and communication by robots. Statements communicated to the robot are mapped to RDF representations, which are stored together with their source. The model also stores the perspective of the source on a property expressed in the statement. The possible perspective values are denial/confirmation, sentiment/emotion, and certainty. One of the purposes of modeling the perspectives is to respond to conflicting or uncertain

interpretations of signals by the model: negative feedback, correct or provide additional information.

In Table 1, we show a simplified example of an RDF representation in the brain which is the result of processing an utterance in a chat for which *Tom* is the speaker, within a specific context in Armando's office on the 24th of January 2019 during which she also perceived a chair and a person. *Tom* claimed that *Karla lived in Paris* and expressed a perspective: he confirms the claim and he is certain and surprised. In the meantime, while *L* was listening to *Tom*, she also saw a chair and recognized a person, *Gabriela* in the room where the chat took place, *Armando's office*. The event and the perceptions are all part of the same context, anchored in time and place. The RDF representation gives further details on the source, as well as perspective, entities and relations expressed in the utterance.

Table 1. RDF triples representing a context taking place in a specific time and place, an utterance in a chat, the speaker, the claim made and the perspective of the speaker on the claim

Named graph: lTalk:Interactions		
lContext:context1	a	eps:Context;
	sem:hasBeginTimeStamp	lContext:2019-01-24;
	sem:hasPlace	lContext:armandosOffice;
	sem:hasEvent	lTalk:chat4;
	eps:hasDetection	lWorld:gabriela, lWorld:chair1
lTalk:chat4	a	grasp:Chat;
	sem:hasSubevent	lTalk:chat4_utterance1
lTalk:chat4_utterance1	a	grasp:Utterance;
	sem:hasActor	lFriends:tom
lContext:armandosOffice	a	sem:Place
lFriends:tom	a	sem:Actor, grasp:Source
Named graph: lTalk:Perspectives		
lTalk:chat4_utterance1_char0-25	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris;
	prov:wasDerivedFrom	lTalk:chat4_utterance1;
	prov:wasAttributedTo	lFriends:tom
lTalk:chat4_utterance1_char0-25_ATTR1	a	grasp:Attribution;
	rdf:value	grasp:CONFIRM, grasp:CERTAIN,
		grasp:SURPRISE;
	grasp:isAttributionFor	lTalk:chat4_utterance1_char0-25
Named graph: lWorld:Instances		
lWorld:karla	a	n2mu:Person, gaf:Instance
lWorld:paris	a	n2mu:Location, gaf:Instance
lWorld:gabriela	a	n2mu:Person, gaf:Instance
lWorld:chair1	a	n2mu:object, gaf:Instance
Named graph: lWorld:Claims		
lWorld:karla_livedIn_paris	a	grasp:Statement, sem:Event
Named graph: lWorld:karla_livedIn_paris		
lWorld:karla	lWorld:livedIn	lWorld:paris

Communication modeling starts with representing the `Context`, which provides information about the situation within which conversations take place. Within a `Context`, there are `Chats`, which model one-to-one conversation. Within a `Chat`, `Utterance`s are spoken, both by the human and the robot. These `Utterance`s are parsed, to obtain a subject-predicate-object RDF `Triple`. The parsed `Utterance` is sent to the brain (represented as in Table 1), which, in response, produces `Thoughts`. These `Thoughts` are the result of the inclusion of the new RDF triple and its reasoning in relation to all stored knowledge.

Figure 3 shows the types of thoughts that we defined so far: *gaps*, *conflicts*, *overlap* and *novelty*. `Gaps` are defined by the ontologies included, and as such relate to the structure of the modelled world. `Conflicts`, `Overlaps` and `Novelty` are defined by the stored triples and relate to the population of the modelled world. A detailed description of the thoughts is given in Table 2. Each of these thoughts represents a state of the brain that requires a communicative action from the robot to improve this state or to inform friends.

4 Context Implementation

In order to provide our semantic agent with sufficient contextual information to properly interpret the world, we split the contextual model into three complementary aspects. Firstly, *discourse* context is modeled as the awareness of what has previously been said. Discourse context is typically distinguished from the *perceptual* or *situational* context. Situational context includes the relevant aspects of the environment, which is non-textual in nature. It consists of date, time, geo-location, and objects and persons present in the space. This allows for a basic spatial awareness of the robot's surroundings.

However, context should not be thought of as a static backdrop to language. Rather, it is dynamic in nature as context not only influences the language used on particular occasions, but is also itself changed by language. Hence, we can perceive an interactive relationship between language and context.

Table 2. Types of thoughts

Cardinality conflict	Claims that cannot coexist because a strict one to one predicate is enforced, but two different objects have been linked
Negation Conflict	A new claim is directly negated by a previous claim
Statement Novelty	Awareness that knowledge was acquired before, along with the provenance, or if it represents genuinely new information
Entity Novelty	Awareness that a new entity is mentioned
Subject/Object Gap	Potential knowledge about a subject/object is absent and provides an opportunity to learn something new
Overlap	Awareness that new claim contain shared, but not equal, information already present in the brain
Trust	Score based on how much people talked, how much the robot learned from them, and how many conflicts they generate

We define this dynamic relationship as the *social* context. The social aspect of this dynamic relationship can be partially explained with Gricean maxims [8]. Essentially, talking to someone means we have a better understanding of their world knowledge as we are creating our mutual social context. Within it, we can describe concepts in a way specific to our individual and mutual world knowledge, differently than we would with someone previously unknown. This personal language is evolving and allowing for shorter references and thus more efficient communication.

The form of referring expression used by a speaker signals their belief with respect to the status of the referent within the hearer's set of beliefs. For example, a pronominal reference signals that the intended referent has a high degree of salience within the hearer's current mental model of the discourse context. [9] Thus, references made in dialogue can be interpreted against the current linguistic and situational context, as well as the growing world knowledge which L shares with the speaker.

In the next subsections, we illustrate the role of context for interpretation and communication within our robot model for the discourse, spatial and social context. The technical reality of this interaction is both challenging and confronting with respect to the questions and problems to be solved for robot-people interaction, setting a research agenda for the future.

4.1 Discourse Context

In discourse models, linguistic context is perceived as a part of general, situational context. This means that within one *situation* there can be many dialogues with different people. Due to the lack of mobility while being in active mode and difficulties of conducting dialogue with more than one person, this world view is well suited for our agent.

Discourse context is stored in a hierarchy of objects which connect the information about the speaker with the current spatial and linguistic context.

Spatial and situational context awareness is modeled with a wide-scope *Context* object. Every context has a unique ID which points to a location such as *Piek's office*. The basic assumption is that people and objects that are present may change but the location and situation may stay the same. Accordingly over time, L will be able to reason over patterns of objects and people present at different contexts, and learn what to expect. A crucial aspect is the identity of the human discourse participant, which is established by face recognition and getting to know new people.

When a conversation starts with an identified person, a new *Chat* object is created which is connected to the speaker. This way, all first and second person personal pronouns can be co-referenced. Within a chat, there can be many *Utterances*, both statements and questions. Keeping track of the types of the things mentioned in recent discourse allows for easier entity and pronoun coreference. For example consider the following utterance: "My sister and I like London, but we've never been there". After syntactic parsing, pronouns are dereferenced using a lexicon and a rule-based system. Coreferring *there* to *London* can be done

simply by knowing that *London* is a location and that *there* is a non-ambiguous pronoun, referring to a location. However, in order to connect the other two co-referents, a slightly higher awareness is necessary. Of course, *I* refers to first person singular, and *we* to first person plural, but *my sister and I* refers to two people, which can then be co-referred with *we*, of which *I* refers to the identified speaker and *sister* is the individual identified relative to the speaker.

All aspects of this model suffer from practical limitations, and the model of linguistic context is no different. In a pipeline system such as this one, there are a lot of possible reasons why the input could not be correctly parsed or interpreted. The input sentence could be ungrammatical or incorrectly transcribed, the Part-of-Speech tagger or the syntactic parser could fail, semantic types can be incorrectly classified, etc. In these cases, a good property of linguistic context is that it is dynamic and the agent can change it too. In other words, by implementing a multi-initiated discourse model, our agent asks for clarification and more information when needed. Confusion, conflicts and uncertainty of coreference relations will thus trigger a drive to resolve these and trigger the robot to ask questions to the human.

4.2 Spatial Context

One of the major problems for our robot is distinguishing between separate instances of objects of the same type. Whereas people are identified individually through face recognition, object recognition only yields types. In the first version of our model, only a single instance of each object type is represented in the brain and all knowledge is linked to this instance, i.e. all perceived chairs result in the same object instance of the type chair: one-type-one-instance. We thus create a single URI for a unique instance based on the type, as an instance of this type for all encounters and mentions. The alternative extreme is to treat each perception and mention of an object type as a new instance, but that overgenerates instances, i.e. one-perception/mention-one-instance. In that case, we create a single URI for each perception, as different instances of the type. The proper granularity of identities is somewhere in between those two extremes but needs to be carefully crafted in accordance with the human ways of defining instances in context.

Failing to distinguish objects (and also people) results in unwanted errors and conflicts, as all claims made about any chair are stored as claims for the same chair. Failing to identify objects results in dispersed information over false identities and more ambiguity, making it impossible to decide which chair is being referenced or e.g. which laptop is my laptop. How then to define the permanence of objects and their identity, so that we approximate the true number of distinct objects per situation?

Our current solution exploits the knowledge about locations and contexts to reason over object instances. As explained in Sect. 3, situations encountered by *L* are represented as instances of a *context*. An instance of a context is anchored in time and connected to a location. All objects and people that she meets during a context instance are linked to this context together with the identified location.

Identifying the location and identifying the objects mutually depend on each other and this forms the basis for making reference to identities in a context.

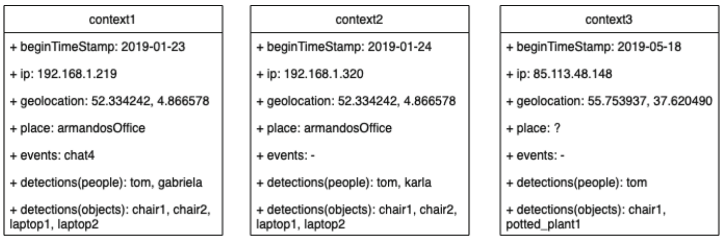

Fig. 4. Example for context construction, and location and object identity

In practice, when switched on, the robot becomes aware of a new context and creates a new instance in her brain. This is shown in Fig. 4, for *context1*, *context2* and *context3* which are created on different days during which she is switched on. Next, she scans the objects and people in her environment and relates them to this new context. People are identified through face recognition and objects are represented as potential new object instances of a certain type based on image recognition. After this first scan, the robot tries to identify her location for which she gathers some initial information (IP, geolocation). She matches all the information of the current context with all previously modeled contexts.

In Fig. 4, the information collected for context2 is compared to context1, whereas context3 will be compared to context2 and context1. Note that only so-called *endurants*, as defined in DOLCE [13], make sense to compare. Endurants, such as object and physical places, persist through time and place and therefore across contexts, whereas perdurants, such as events, conversations, and situations only exist within a time and place boundary and therefore only exist at most for the duration of each instance of a context. Given the basic information on the location derived from the system, the robot thus only uses physical objects and dimensions to compare contexts for determining the potential location.

If there is sufficient overlap with a previous context, L hypothesizes that she is now in the same location. In case of uncertainty, she can ask for confirmation. If she is certain that there is no match, she assumes she is in a new location and will ask for its name. If a new location is detected and confirmed, the robot assumes all objects in this location are new instances. If a known location is recognized, she will map the physical objects of the new context to the objects of the matched location of the most recent context. If there are fewer objects in the new context, these objects are assumed to be absent but still exist in the brain. If there are more objects in the new context, new instances are created to match the cardinality. Object identity is thus determined in relation to location identity, where the robot tries to maximize the permanence of objects for each

location across different contexts. Obviously, we need to extend this model to deal with objects that can move to new locations.[2]

In Fig. 4 for example, context2 matches context1 for *Tom* and two chairs and two laptops. On the basis of the match, *L* concludes she is now in *amandosOffice* and the chairs and laptops are assumed to be the same, as there is no cardinality mismatch. What is different is the presence of *Gabriela* in context1 and the presence of *Karla* in context2. In contrast in the case of context3, only *Tom* and one *chair* are matched while the *potted_plant* is new. Therefore, the place value remains unresolved which will trigger her to ask for the location. If that is different from previous locations, both the *chair* and the *potted_plant* will be added as new instances to the brain, each with properties indicating when and where they were perceived and mentioned if referenced during a chat in the same context.

In communication, the robot treats objects in new locations as new instances unless told otherwise. For example, if somebody claims ownership of a chair within a context and location, e.g. *this is my chair*, the property *own* is assigned to that instance. In another location, a similar object can be perceived but it is considered to be a different instance. However, if the same person again claims ownership of this similar object, the robot realizes that multiple similar objects related to different locations are owned by the same person. As a weak conflict, this may trigger questions about identity: *is this the same chair?* On the other hand, if the chair in this new location is claimed to be owned by another person, it does not result in a conflict as it was already represented as a different chair in the brain and both chairs can have different owners.

Our current implementation can only detect a limited range of objects of coarse types and we have only started to detect basic object properties such as colour, size and relative position. The robot awareness of contexts, locations and the objects is therefore extremely shallow and limited compared to human representations. However, our model is open to more fine-grained representations and improvements in detecting differences. If image recognition improves, future versions can detect even more object properties than colour or specific positions in a location using 3-D triangulation. Likewise, we expect that e.g. ownership of very similar objects can be hypothesized from closeness to the owner: my chair, phone and laptop are close to me, yours are close to you.

4.3 Social Context

Social context is on the one hand defined by the drives for social interaction and on the other hand by the shared personal experience built up from previous encounters. The modeling of knowledge through GRaSP enables the robot to consider all the knowledge, experiences and communication that is the result of encounters with a single specific person. This shared social knowledge represents a personalized context, which forms the basis for more efficient communication.

[2] In the future, we plan to use properties of objects (both perceived and communicated) to help to further separate different instances, e.g. *green chair* or *my chair is close by me*.

1. less ambiguity: both lexical and referential ambiguity is limited as the shared vocabulary and shared world take priority over the possible world in interpretation
2. less variation: words used previously to make reference are preferred over alternative variants
3. more relevance: situations, objects, people, concepts previously discussed are more relevant than others and new information in relation to the known is also more relevant

Following the Gricean Maxims, the shared knowledge and experience defines information that does not need to be exchanged and it also provides that background against which all new signals are to be interpreted. The first time I talk about *my sister* the robot can only identify her relative to me but except for that she has no idea about the identity. As a result of my statement a URI will be created to represent her in the brain as an instance of a person (gender female) and the kinship relation is created to link her to me and my parents. As *L* does not know her name, she will ask for it, which is *Selene*. From now, we share this knowledge and I can refer to her as *my sister* or as *Selene* as shared knowledge.

On the other hand, *Selene* is also the name of another friend of *L* . By introducing my sister, an ambiguity is created. Of course, there are many Selenes in the world, but in my case the ambiguity only exists if I also know the other Selene. References to *Selene* by me are not ambiguous as long as *L* derives from her brain that we only talked about one Selene. On the other hand, she may be prompted to ask if I happen to know the other Selene as well. In that case, two Selenes become part of our shared knowledge and from that point on mentions need to be disambiguated, i.e. *Selene, your sister* and *Selene, my friend*.

The drives that the robot has to pro-actively interact can be tuned to such personalized contexts. Drives such as Novelty, Conflicts, Uncertainty, Trust are mainly considered in relation to you. In case of *Selene*, we already saw that Novelty of information that results from our conversation triggered *L* to inform me about related knowledge she has. Similarly, she may ask if I know people she knows that have any other background that is similar. There can be conflicts and uncertainty coming from the communication with anybody but she will only address me about conflicts and uncertainty that related to what we talked about. Trust is a judgment based on the information I communicated in the past and therefore is personal by definition but also relates to others because it reflects the number of conflicts I am involved in. Finally, even perception is based on our social relationship. She may constantly perceive objects but she will give priority to objects we talked about before or the ones that are close to me or that I own.

5 Related Work

Mavridis [14] gives an overview of natural language processing technologies in human-robot interaction and challenges to be tackled, including 'theory of mind', open-domain communication, varied speech acts, symbol grounding and

multiple-turn dialogues. Most human-robot communication models still only handle basic communication using one or two speech acts, limited symbol grounding and single turns. In fact, none of these systems exploits context.

Embodied dialogue agents nevertheless offer many new challenges and possibilities to exploit the multimodal nature of situated dialogue. When the embodied agent has a conversation with a partner, they interact within a situational context. This encompasses the physical world around them, their positions within it, as well as the moment in time and other pragmatic notions that arise from the situation. This affects the referring expressions used within dialogue, and fluent use of these expressions is affected by the mutual knowledge that the conversational partners share [4]. The choice of referring expressions used is affected by their salience, whether in discourse or within the situational context. In most computational models, the choice of possible referents can be found within the Discourse Context, which is accumulated through conversation [3]. Discourse context is commonly differentiated from the mutual knowledge set, which contains information available to both conversational partners that was not referred to in the discourse. A referring expression in an utterance introduces a representation into the semantics of that utterance and this representation must be bound to an entity in the mutual knowledge set (in the case of evoking or exophoric references) or in the discourse context (in the case of anaphoric references) for the utterance to be resolved [10].

To interpret a referring expression, algorithms typically analyze the recently mentioned entities. However, this is not enough for embodied agents which are becoming more and more prominent in a variety of domains. One kind of embodied agents are robots with an integrated spoken interface. Implementations of such agents commonly focus on command-and-control interfaces, rather than placing the user and conversational agent into a shared space which can be talked about in an open dialogue [3].

A common approach used to systematically represent situations for the purpose of modeling situational context are ontologies. The design of ontologies for this purpose needs to comply with semantic requirements regarding the capabilities of representing contexts and situations in a general way. For instance, the Situational Context Ontology combines contextual information (spatial and temporal) with related situations of an individual [11].

Our model is designed for open communication with the explicit result of acquiring knowledge and building a social relationship. Ontological knowledge is used to control the interaction and to interpret the context, e.g. people, friends, locations, space, objects detected by the image recognition, some basic object properties. This basic ontology allows us to model the context of the interaction and the references to these contexts in the communication. Although the ontological model is relatively basic, it allows us to model and experiment the interaction in real-world situations. We defined a context as an episodic element that explicitly gathers everything *Leolani* learns in connection with specific situations. In addition, we defined *thoughts* as reflections on the interpretations of the contexts and any previous episodic encounter, i.e. *awareness* of gaps,

relevance, conflicts, uncertainties, trust. Reflections result in drives to interact with the participants or observe the situation, which results in an update of the context interpretation.

6 Conclusion

In this paper, we described a robot model for social communication within contexts that was implemented and can be used for further experiments. We did not take a theoretical perspective but tried to consider all practical aspects from a pragmatic perspective dealing with the full complexity of a real-world physical context. By considering the problems within realistic situations, we present a vision for future research in essential but also down-to-earth aspects of interpreting contexts.

We explained three notions of context: spatial, discourse and social, that interact with each other and with the interpretation of signals. We demonstrated that also the context needs to be interpreted as a collection of signals and that contexts and signals define each other. We showed how our models try to exploit this relationship and what the limitations are. The level of mutual understanding of the context and the signals within is still very limited and our robot still has the capacity of less than a child. Partially, these limitations can be resolved by better image recognition (objects, properties and relations), detection of scenarios, more knowledge acquired over longer periods of time, richer language models, and more. The pioneering work described in this paper, sets an agenda to further experiment with the different aspects of context and interpretation in real-world physical environments and to evaluate different model implementations.

Acknowledgement. This project was funded through the NWO-Spinoza funds awarded to Piek Vossen and by the VU University of Amsterdam. We specifically thanks Selene Kolman and Bob van Graft for their support.

References

1. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science (2016)
2. Project authors, T.W.: Webrtc. In: Online publication (2011). https://webrtc.org/
3. Byron, D.: Understanding referring expressions in situated language: some challenges for real-world agents. In: Proceedings of the First International Workshop on Language Understanding and Agents for the Real World, pp. 80–87. Citeseer (2003)
4. Clark, H.H., Marshall, C.R.: Definite reference and mutual knowledge. In: Psycholinguistics: Critical Concepts in Psychology, vol. 414 (2002)
5. Fokkens, A., Vossen, P., Rospocher, M., Hoekstra, R., van Hage, W.: Grasp: grounded representation and source perspective. In: Proceedings of KnowRSH, RANLP-2017 workshop, Varna, Bulgaria (2017)
6. Frege, G.: Über sinn und bedeutung. Reclam, Philipp (2019)

7. Google: Cloud speech-to-text - speech recognition. In: Online publication (2018). https://cloud.google.com/speech-to-text/
8. Grice, H.P.: Logic and conversation. 1975, pp. 41–58 (1975)
9. Kelleher, J.D., Dobnik, S.: Referring to the recently seen: reference and perceptual memory in situated dialog (2019)
10. Kelleher, J.D., Dobnik, S.: Referring to the recently seen: reference and perceptual memory in situated dialog. arXiv preprint arXiv:1903.09866 (2019)
11. Kolbe, N., Zaslavsky, A., Kubler, S., Robert, J., Le Traon, Y.: Enriching a situation awareness framework for IoT with knowledge base and reasoning components. In: Brézillon, P., Turner, R., Penco, C. (eds.) CONTEXT 2017. LNCS (LNAI), vol. 10257, pp. 41–54. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57837-8_4
12. Leslie, A.: Pretense and representation: The origins of "theory of mind". Psychol. Rev. **94**(4), 412 (1987)
13. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Wonderweb deliverable d17. Comput. Sci. Prepr. Arch. **2002**(11), 74–110 (2002)
14. Mavridis, N.: A review of verbal and non-verbal human-robot interactive communication. Robot. Auton. Syst. **63**, 22–35 (2015)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
16. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
17. Szegedy, C., et al.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015). http://arxiv.org/abs/1409.4842
18. Vossen, P., Baez, S., Bajčetić, L., Bašić, S., Kraaijeveld, B.: A communicative robot to learn about us and the world. In: Proceedings of Dialogue-2019, Moscow (2019)
19. Vossen, P., Baez, S., Bajčetić, L., Kraaijeveld, B.: Leolani: a reference machine with a theory of mind for social communication. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2018. LNCS (LNAI), vol. 11107, pp. 15–25. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00794-2_2