A communicative robot to learn about us and the world

Vossen, Piek; Baez Santamaria, Selene; Bajceti, Lenka; Baši, Suzana; Kraaijeveld, Bram

published in
Computational Linguistics and Intellectual Technologies
2019

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

Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

A COMMUNICATIVE ROBOT TO LEARN ABOUT US AND THE WORLD

Vossen P. (piek.vossen@vu.nl),
Baez S. (selene.baez.santamaria@gmail.com),
Bajcetić L. (lenka.bajcetic@gmail.com),
Basić S. (suz.basic@gmail.com),
Kraaijeveld B. (bram.kraaijeveld@gmail.com)
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

We describe a model for a robot that learns about the world and her companions through natural language communication. The model supports open-domain learning, where the robot has a drive to learn about new concepts, new friends, and new properties of friends and concept instances. The robot tries to fill gaps, resolve uncertainties and resolve conflicts. The absorbed knowledge consists of everything people tell her, the situations and objects she perceives and whatever she finds on the web. The results of her interactions and perceptions are kept in an RDF triple store to enable reasoning over her knowledge and experiences. The robot uses a theory of mind to keep track of who said what, when and where. Accumulating knowledge results in complex states to which the robot needs to respond. In this paper, we look into two specific aspects of such complex knowledge states: 1) reflecting on the status of the knowledge acquired through a new notion of thoughts and 2) defining the context during which knowledge is acquired. Thoughts form the basis for drives on which the robot communicates. We capture episodic contexts to keep instances of objects apart across different locations, which results in differentiating the acquired knowledge over specific encounters. Both aspects make the communication more dynamic and result in more initiatives by the robot.

Keywords: multimodal communication, social robots, knowledge acquisition and modeling

1. Introduction

Human-robot communication is necessary for collaboration in future societies. It is vital to build social relationships between humans and robots, to create a common ground from shared experiences and knowledge, and to build up trust. Natural language communication in multimodal environments plays a crucial role for establishing such a relationship.

Both machines and humans make errors in dealing with real-life situations. We have therefore designed a robot model that assumes that information can be wrong, has gaps and even conflicts. To deal with this, the robot needs to learn about us and the world: fill gaps and get feedback on errors and confirmation in case of uncertainty. In previous work, [17], we described a female robot model, named *Leolani*, that supports open-domain learning through communication, having a drive to learn new concepts and make new friends. The absorbed knowledge consists of everything people tell her, the situations and objects she perceives, and what she finds on the web. The results of her interactions and perceptions are kept in a triple store, enabling her to reason over her knowledge and experiences. The robot uses a theory of mind [7] to record the learning provenance (who said what, when and where).

Learning through communication results in complex knowledge states that may contain errors, false statements, conflicts or interpretations that differ across different people and situations. The functioning of the robot is at risk if the acquired information is taken as it is. It is therefore necessary that the robot knows how to reflect on the state of her brain and takes initiatives to improve this state. Furthermore, situations need to be interpreted within the unique context of an interaction. Knowledge that is accumulated within such a situation needs to be related to this context as well, e.g. my laptop is likely to be found in my office but not in other places. By differentiating these contexts, possible conflicts can be prevented and communication will be easier as there is less ambiguity and fewer conflicts.

In this position paper, we therefore describe an extension to *Leolani* that reflects on the acquired knowledge by producing so-called *thoughts*. These thoughts result in drives to improve the state of brain through communication. The robot takes initiatives to involve her human sources for that purpose. The robot model also includes a notion of context that allows us to identify different situations and the objects within it. This results in fewer conflicts and less confusion (uncertainty) and therefore more healthy states of the brain, better definitions of relevance and less need to communicate.

This paper is structured as follows: In **Section 2**, we summarize related work on social robot communication. Our data model and the way in which the robot learns through communication are described in **Section 3**. In **Section 4**, we describe the thoughts and the corresponding drives that lead to initiatives to communicate. For dealing with the world and humans, the robot needs to represent and memorize the contexts in which she encounters people. In **Section 5**), we explain how instances of contexts are created and how these result in more fine-grained and differentiated representations of situations. We conclude and discuss future work in **Section 6**.

2. Related work

Mavridis [11] 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.

Recently, there has been an increase in chat systems that can be used for human-robot communication. Many of these models are either scripted ([14],[1]) or based on neural networks (often sequence-to-sequence (seq2seq) models), see for example: the dialogue systems built from the Ubuntu dialogue corpus [9], CoQA corpus [12], Twitter [8], the Persona-Chat dataset [18] and movie dialogues ([13] and [16]). Both types can be seen as extremes on the scales of control and fluency. Scripted conversations allow developers to control interaction, but knowledge needs to be defined manually and the conversation is limited, not robust and rarely fluent. Seq2seq models, on the other hand, are robust, fluent and respond to any input, but cannot be controlled or explained. More importantly, no explicit knowledge is derived from these conversations.

Our model is designed for open communication with the explicit result of acquiring knowledge and building a social relationship. It is designed for generic purposes defined at a low level that can support any high-level goal. This architecture provides our model with more flexibility and fluency than strictly scripted models, while the communication is more purposeful than in seq2seq models.

Another important aspect of human-robot communication is mixed-initiative interaction. Many systems leave the initiative to the human and only respond when prompted. They do not have an intrinsic drive to communicate unless they are scripted for some task, e.g. to take your order. Little work has been done on the implementation of basic drives to communicate in the systems. Our model implements low-level drives, such as the need to fill knowledge gaps and resolve conflicts and uncertainty. These drives make the communication active, lively and purposeful. We do not intend the model to fully capture human dialogue. Rather, dialogues serve to satisfy the robot's drives.

In our previous paper [17], we focused on a robot with a theory of mind [7] that acquires knowledge from people but stores the knowledge as claims from these people. In this paper we add the notions of *thought* and *context*. A thought represents a brain state that triggers drives. A context is an episodic element that explicitly gathers everything *Leolani* learns in connection with specific situations. *Thoughts* and *context* pave the way for new cognitive functionalities like relevance and permanence, as well as new intentions that exploit contextual information to drive the conversation. They also equipt the robot with new initiatives for communication and at the same time reduce conflicts, ambiguities and define relevance.

3. Data Model

3.1. Model description

Our robot model architecture is shown in Figure 1. We defined four layers:
1) a sensor processing layer,
2) a communication layer that responds to sensor input or inner drives,
3) a language processing layer to deal with questions and statements, and
4) a 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 [3] for speech detection, the Inception neural network [15] for object recognition, OpenFace [2] for face recognition, and Google Cloud Speech-to-Text API [5] for speech recognition. We use the outputs of these processing modules as inputs to the other layers. Therefore, we do not address potential conflicts and ambiguities in the signal layer itself, but try to resolve them in the higher-level layers.

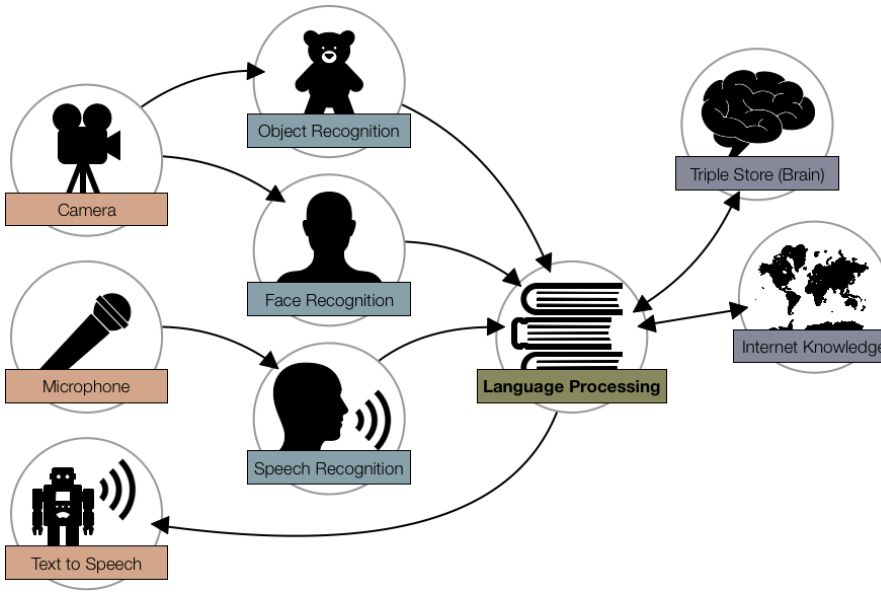

Figure 1: Global architecture of the robot model

In this paper, we focus on modeling the result of communication in an RDF triple store (called 'the brain'), which stores all interpretations of experiences. The brain forms the basis for the drives of the robot to communicate. We use the Grounded Representation and Source Perspective (GRaSP) model [4] 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 the source of each statement.

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. Besides processing statements, the robot handles questions as SPARQL queries against the knowledge in the brain.

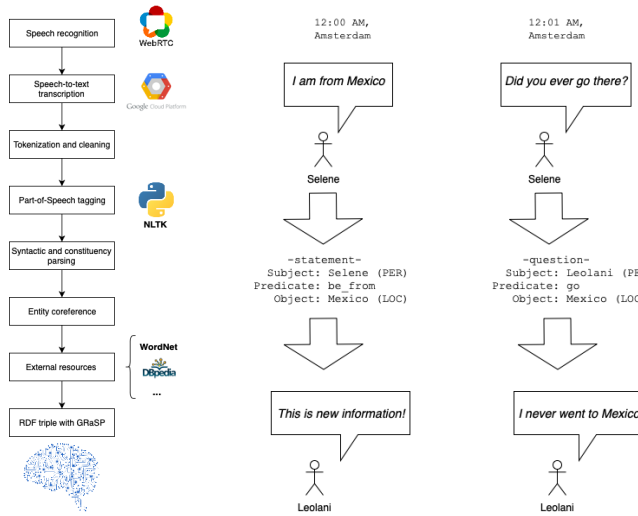

Figure 2: Natural Language Processing Pipeline

As shown in Figure 2, the NLP Pipeline consists of several external components, while some are manually implemented specifically for this task. For the sake of transparency, we resorted to rule-based parsing instead of a neural-net approach. This refers specifically to the syntactic and constituency parsing. Syntactic parsing is done with a Context-Free Grammar which captures the most typical sentence constructions in English. Since English has quite a strict word order, making such a grammar was manageable. After the CFG grammar creates a tree from the sentence, the tree is passed on to the Constituency parser, which assigns roles to the tree nodes. This is done by relying on word order, but also the POS tags and, if necessary, semantic types. The constituency parser outputs a triple, consisting of a subject, a predicate and an object, which can be stored in the brain as a claim or used to query it. Furthermore, to extract perspective information we rrsort to a simple lexicon of typical sentiment and certainty predicates, such as *like* and *think*. These lexical verbs, along with modal verbs and polarity markers, e.g. *never*, are suited for a rough estimate of the perspective expressed by the speaker.[1]

In Table 1, we show a simplified 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

[1] As a next step, the model will include temporality within the perspective, using a lexicon of temporal expressions and a more advanced morphological analysis of predicate tense. Temporality indicates whether the statement is about the here and now, the past or the future (irrealis)

a perspective: he confirms the claim and he is certain and surprised. In the meantime, while *Leolani* 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 that is anchored in time and place. The RDF representation gives further details on the source and the perspective and the entities and relations expressed in the claim.

Table 1: RDF representation 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;
	eps:hasDetection	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	ATTR1a	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.

3.2. Model implementation

Following the model in Figure 1, the robot world is implemented both as a Python application, shown in Figure 3, and as an RDF representation, shown in Figure 4.

Communication modeling starts with representing the `Context`, which provides information about the situation within which conversations take place. Within

a **`Context`**, there are **`Chat`**s, which model human-robot one-to-one conversation. Within a **`Chat`**, **`Utterance`**s are spoken, both by the human and the robot. These **`Utterance`**s are parsed, as mentioned in Section 3.1, 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.

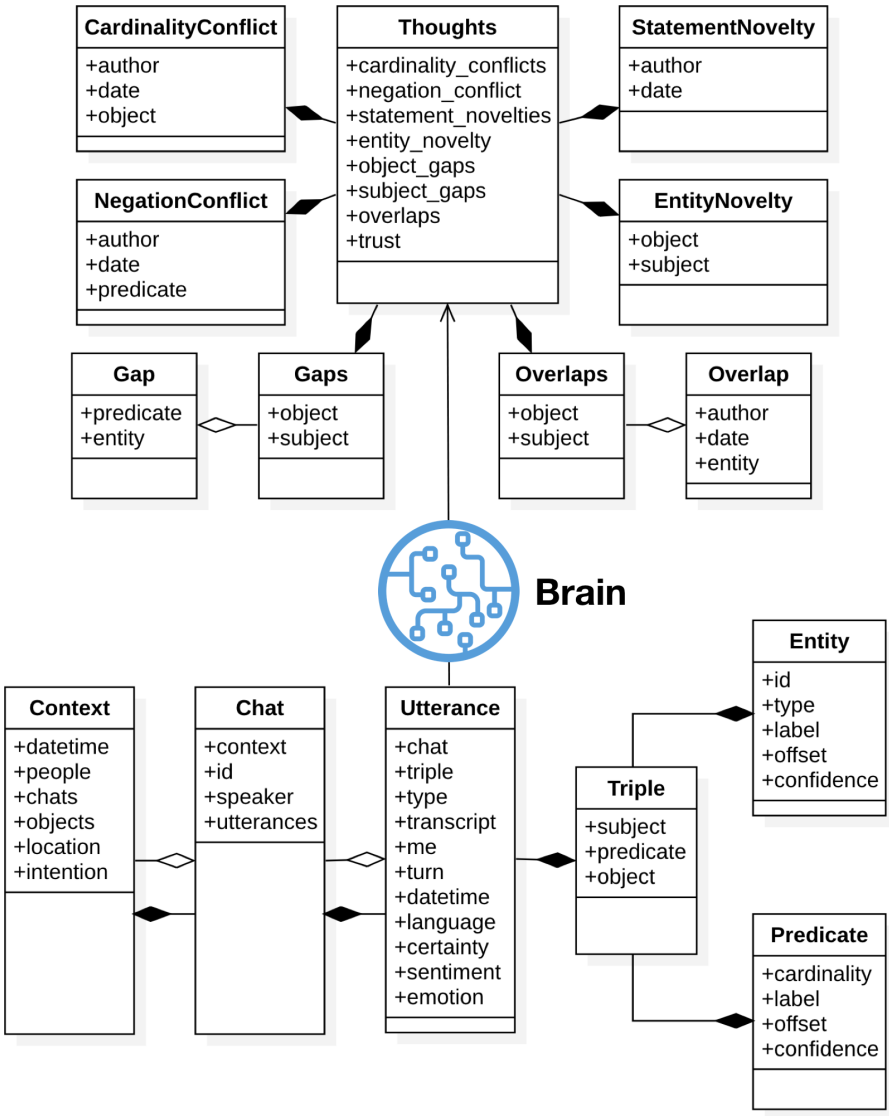

Figure 3: Data model class diagram

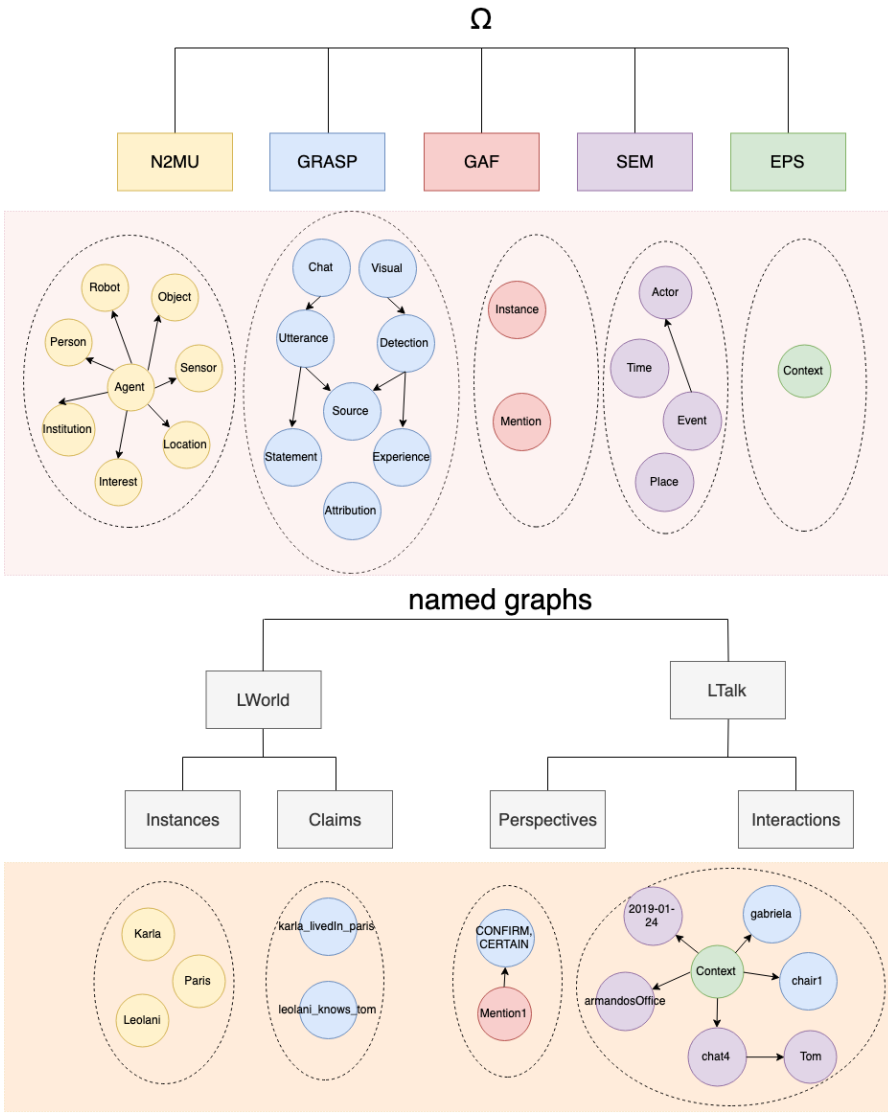

Figure 4: RDF representation

Figure 3 shows the different 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 content of the modelled world. A detailed description of what these thoughts represent is presented in Table 2. Each of these thoughts represents a state of the brain that requires a communicative action from the robot which is implemented as a drive, for example to improve this state or to inform friends. The way these **`Thoughts`** generate drives is explained in the next section.

Table 2: Types of thoughts

Cardinality Conflict	statements that cannot coexist because only one object is allowed
Negation Conflict	a previous statement is directly negated by a person
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 Gap	potential knowledge about a subject is absent and provides an opportunity to learn something new
Object Gap	potential knowledge about an object is absent and provides an opportunity to learn something new
Overlap	awareness that new statements contain shared, but not equal, information already present in the brain
Trust	a score based on how much people talked, how much the robot learned from them, and how many conflicts they generate

4. Drives

In passive robot models, people ask questions or make statements to which a robot responds. However, it may prove useful to equip a robot with drives to optimize its relation with humans and to learn from interactions. In a high-level task, e.g. finding and moving objects or showing the way, the robot can take initiative to achieve the goal. In our current model, we focus on lower-level drives that can play a role within any high-level task. Here, we specifically focus on two tasks to explain the notion of drives: 1) open-ended learning and 2) creating a personal relationship involving shared knowledge, experiences and trust. Next, we discuss some drives and thoughts related to these tasks and the corresponding communication in more detail.

4.1. Getting to know people

Knowing people is one of the robot's primary drives, as they are important sources of knowledge. The robot keeps track of her human sources through face recognition. When she meets a new person, she is triggered to learn about this person. This trigger is the result of a **`SubjectGap`**. The properties asked are predefined by the Nice2-MeetYou (n2mu) ontology, which captures social properties to start the communication, e.g. where are you from, what you like and who you know. For example, after meeting *Karla*, the triples in Table 3 inform *Leolani* that she does not know where Karla lives or what her favorite interest is.

Table 3: Sample supporting triples to infer a SubjectGap

lWorld:Karla	a	n2mu:Person .
n2mu:Person	livedIn	n2mu:City .
n2mu:Person	favorite	n2mu:Interest .

After learning about a new person, the robot queries the brain to check if other people have a similar property. An **Overlap** thought is generated if the new statement contains some shared, but not equal, information already present in the brain. For example, the triples in Table 4 show that "Karla lives in Paris" would generate an overlap with "Tom lives in Paris". The resulting **Overlap** prompts her to respond *Do you know my friend Tom who also lives in Paris?*.

Table 4: Sample supporting triples to infer an Overlap

lWorld:Karla	livedIn	lWorld:Paris .
lWorld:Tom	livedIn	lWorld:Paris .

4.2. Open-ended learning from conversation

In the above example, learning is driven by the predefined ontology. The ontology defines the properties as in a closed world, e.g. *like*, *know*, *origin*, *own*. However, we do not predefine the objects of these properties. Statements such as *I like Scrappy_Doo* or Tom likes Felix are taken seriously and the object is always stored as in instance labeled by the text coming from the speech recognition without further interpretation.

If an object is not defined in the brain by at least the type of thing it is, an **ObjectGap** thought it derived which triggers the robot to learn about it. She either asks people or consults the web. Asking people *What is Scrappy_Doo?*, she may learn it is a dog. Consulting the web what a dog is, she may learn that a *dog* is a *mammal* according to DBpedia. Asking people what a dog is, she may learn it is a *pet*. Learning about objects, can result in further thoughts such as **Overlap**, which may yield again other triggers. For example, Table 5 reflects that *Leolani* can infer that *Karla likes dogs* because she learned that *Scrappy_Doo* is a dog and *Karla likes Scrappy_Doo*. Learning that dogs are mammals may make her think that *Karla like mammals*. Knowing that cats are also mammals she can hypothesize that *Karla likes cats* and even that Karla may like *Felix*. This may make her ask Karla *Do you like Felix too?*.

Table 5: Sample supporting triples to infer a ObjectGap

lWorld:Karla	n2mu:like	lWorld:Scrappy_Doo .
dbr:Scrappy-Doo	dbo:species	dbr:Dog .
dbr:Dog	a	dbo:Mammal .
lWorld:Tom	n2mu:like	lWorld:Felix_the_cat .
lWorld:Felix_the_cat	a	n2mu:cat, dbo:Mammal .

4.3. Relevance and novelty

StatementNovelty determines if *Leolani* has acquired this knowledge before, along with the provenance information, e.g. when Karla states "I lived in Paris", *Leolani* can identify that she has heard this before from Tom. This may trigger informing Karla about this. **EntityNovelty** also signals if the statement involves a new entity, either

as the triple's subject or object. For example, "Karla visited Morocco" could lead to *Leolani* realizing she never heard about Morocco before. In general, *Leolani* comments on novelty to her friends, telling them what she learned: these are **`StatementNovelty`** thoughts.

Table 6: Sample supporting triples to infer a StatementNovelty

lTalk:chat4_utterance1_ char0-25	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris ;
	prov:wasDerivedFrom	lTalk:chat4_utterance1 ;
	prov:wasAttributedTo	lFriends:tom .
lTalk:chat5_utterance1_ char0-16	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris ;
	prov:wasDerivedFrom	lTalk:chat5_utterance1 ;
	prov:wasAttributedTo	lFriends:karla .

`Novelty` and **`Gap`** thoughts also yield a risk: the robot may continue talking and asking questions forever to learn more. She lacks Gricean maxims of relevance and quantity [6]. We currently limit such drives by randomly selecting responses if there are too many and mimicking relevance through recency and relatedness to the speaker. New information about the currently addressed person is considered highly relevant. Similarly, new information connecting to knowledge previously discussed with the addressee is relevant. In any case, recent information is more urgent and relevant than old information.

4.4. Uncertainties, conflicts and ambiguities

Open-ended learning also entails a risk with respect to information quality. We currently address this by capturing uncertainty scores for knowledge and perceptions, by detecting conflicts and by resolving ambiguities. Table 7 shows some of the uncertainties *Leolani* encounters.

Table 7: Types of uncertainty. * represents future work

The identity of the human participant	confidence scores of face detection confidence scores of name detection
Ambiguity in language	guessing based on immediate context
Object detection	confidence of the type mismatch with previous encounters*
Speech detection	confidence scores from the speech level of noise in the environment*
Uncertainty expressed by the human participant	classifiers that detect modal expressions classifiers that detect uncertainty from the speech itself: corrections, hesitations, volume* number of corrections, negative feedback*

The types of conflicts currently modeled are **`CardinalityConflict`**s and **`NegationConflict`**s. *Leolani* immediately addresses the source when a conflict arises and confronts other sources that provided the primary information.

A **`CardinalityConflict`** is produced whenever an author claims a statement that can not coexist with another statement as it involves a strictly one-to-one predicate. For instance, "Karla was born in France" cannot coexist with "Karla was born in Japan". A **`NegationConflict`** is returned when an author claims a direct negation of an already learned statement. For instance, "Karla lives in Paris" cannot coexist with "Karla does not live in Paris". These kinds of conflicts trigger *Leolani* to ask people for further clarification.

Table 8: Sample supporting triples to infer a CardinalityConflict

lWorld:Karla	n2mu:bornIn	lWorld:france .
lWorld:Karla	n2mu:bornIn	lWorld:japan .

Table 9: Sample supporting triples to infer a NegationConflict

lTalk:chat4_utterance1_char0-25_ATTR1 a	grasp:Attribution;
rdf:value	grasp:CONFIRM, grasp:CERTAIN, grasp:SURPRISE;
grasp:isAttributionFor	chat4_utterance1_char0-25.
lTalk:chat5_utterance1_char0-16_ATTR2 a	grasp:Attribution;
rdf:value	grasp:DENY, grasp:CERTAIN;
grasp:isAttributionFor	lTalk:chat5_utterance1_char0-16.

In our current implementation, *Leolani* only reports uncertainties and conflicts. Having a theory of mind means that conflicting information does not pose an issue. It is important that conflicting information can be stored and talked about, as this helps *Leolani* function in our conflicting and ambiguous world. In a future version, we implement more specific strategies to resolve them, e.g. consulting other (trustworthy) sources to get confirmation (e.g. DBpedia). Eventually, she could distill her own judgment based on gathered evidence.

Resolving ambiguity that is inherent to natural language is done by keeping track of the linguistic context. For instance, third person pronouns are disambiguated using the information on the last mentioned person and the information on gender. The system is equipped with a lexicon of pronouns, which contains information on the type of entity the pronoun can stand for. Cross-referencing this information with the knowledge of semantic types of previously mentioned entities allows *Leolani* to quickly guess what the pronoun might refer to. Guessing is only done when there is a high certainty level, otherwise *Leolani* will declare her confusion and ask "Which he/she do you mean?". Future plans include expanding the questions to refer to the potential guesses, like this "When you say 'she', do you mean your sister?". By relying on linguistic context and salience, we create a proactive approach to disambiguation and entity coreference, well-suited for a mixed-initiative dialogue system.

4.5. Building trust

The GRaSP model results in the accumulation of claims and the sources of those claims. Over time, the brain provides information about: 1) who shares claims with whom, 2) how many people believe or deny a claim, 3) how certain people are, both generally and individually, 4) how much emotion is expressed by whom, 5) who changes their opinion when and how often, 6) who tells things about others that are denied by the primary source, 7) who has provided most knowledge and how trustworthy that knowledge is, 8) the number of conflicts raised by a source. All this information can be used to build up trust with companions.

At the moment, **`Trust`** involves a score for people she speaks to, based on how much they have talked, how much she has learned from them, and how many conflicts they generate. Furthermore, trust can generate *thoughts* that may trigger new actions or it can be used to respond differently in case of conflicts or uncertainties in a future extension of the model. Information learned from trustworthy speakers is regarded as more likely to be correct.

5. Context awareness

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: all-perceptions-one-instance. The alternative is to treat each perception of an object type as a new instance of said object, but that over-generates instances, i.e. one-perception-one-instance. 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. How then to define the permanence of objects and their identity, so that we achieve a natural balance for representing objects per situation and not too many?

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

This is how it works. When switched on, the robot becomes aware of a new context and creates a new instance in her brain. This is shown in **Figure 5**, 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.

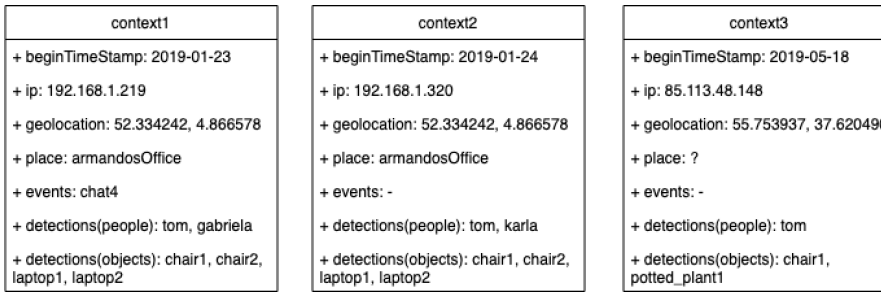

Figure 5: Example for context construction, and location and object identity

In **Figure 5**, the information collected for context2 is compared to context1, whereas context3 will be compared to context2 and context1. Note that only properties with so-called *endurants* as objects make sense to compare. As defined in the DOLCE ontology [10], endurants, such as objects and physical places, persist through time and place, whereas perdurants, such as events, conversations, time 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, *Leolani* 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 less 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.

In **Figure 5** for example, context2 matches context1 for *Tom* and two chairs and two laptops. On the basis of the match, *Leolani* 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.[2] In contrast in the case of context3, only *Tom* and one *chair* are

[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*.

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.

In communication, the robot treats objects in new locations as new instances unless told otherwise. For example, if somebody claims ownership of n 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.

6. Conclusion

In this position paper, we described our models and implementation for a robot that can learn through communication for the purpose of building a social relationship. Our model stores knowledge as triples with the source and its perspective. It represents communication as chats and turns in which claims are made. The model allows the robot to deal with knowledge coming from different sources, handle uncertainties and conflicts, and derive trust in sources. The robot uses thoughts representing states of the brain, which trigger actions and communication as low-level drives. Finally, we have shown how the robot creates an episodic representation of a context linked to time and location, with awareness of the presence of people and objects. Awareness of contexts and locations can be used to identify object instances and model the permanence of objects. All the code of our model is available on Github[3] and project progress is reported on our website[4].

Currently, the robot has acquired knowledge regarding 296 statements through 164 conversations held with 26 distinct people. These conversations were held for testing the system and we have not evaluated the quality. In the future, we plan to carry out experiments to measure the performance of our model. Intrinsic evaluations should demonstrate the capacity to understand humans and the world, to acquire knowledge, to acquire vocabulary and expressions, and to express drives to improve the state of the brain. Extrinsic evaluations should demonstrate the user satisfaction, the quality of the relationships and any high-level task that is modeled. For evaluations, we need to create evaluation data and scenarios, define criteria and create baselines and alternative models.

[3] https://github.com/cltl/pepper

[4] http://makerobotstalk.nl/

Acknowledgements

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. *Abdul-Kader, S. A., Woods, J.:* Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications. 6, 7, (2015).
2. *Amos, B. et al.:* OpenFace: A general-purpose face recognition library with mobile applications. CMU-CS-16-118, CMU School of Computer Science (2016).
3. *T. W. project:* WebRTC. In: Online publication. (2011).
4. *Fokkens, A. et al.:* Grasp: Grounded representation and source perspective. In: Proceedings of knowrsh, ranlp-2017 workshop, varna, bulgaria. (2017).
5. *Google:* Cloud speech-to-text - speech recognition. In: Online publication. (2018).
6. *Grice, H. P.:* Logic and conversation. 1975. 41–58 (1975).
7. *Leslie, A.:* Pretense and representation: The origins of "theory of mind.". Psychological review. 4, (1987).
8. *Li, J. et al.:* A persona-based neural conversation model. arXiv preprint arXiv:1603.06155. (2016).
9. *Lowe, R. et al.:* The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909. (2015).
10. *Masolo, C. et al.:* Wonderweb deliverable d17. Computer Science Preprint Archive. 2002, 11, 74–110 (2002).
11. *Mavridis, N.:* A review of verbal and non-verbal human–robot interactive communication. Robotics and Autonomous Systems. 63, 22–35 (2015).
12. *Reddy, S. et al.:* Coqa: A conversational question answering challenge. arXiv preprint arXiv:1808.07042. (2018).
13. *Serban, I. V. et al.:* Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. pp. 3776–3784 (2016).
14. *Spekman, M. L. et al.:* Perceptions of healthcare robots as a function of emotion-based coping: The importance of coping appraisals and coping strategies. Computers in Human Behavior. 85, 308–318 (2018).
15. *Szegedy, C. et al.:* Going deeper with convolutions. In: Computer vision and pattern recognition (cvpr). (2015).
16. *Vinyals, O., Le, Q.:* A neural conversational model. arXiv preprint arXiv:1506.05869. (2015).
17. *Vossen, P. et al.:* Leolani: A reference machine with a theory of mind for social communication. In: Proceedings of tsd-2018, brno, https://www.tsdconference.org/tsd2018. (2018).
18. *Zhang, S. et al.:* Personalizing dialogue agents: I have a dog, do you have pets too? arXiv preprint arXiv:1801.07243. (2018).