Acquisition of Word-Object Associations from Human-Robot and Human-Human Dialogues

Sepideh Sadeghi[1], Brad Oosterveld[1], Evan Krause[1], Matthias Scheutz[1]

Abstract— **Past work on acquisition of word-object associations in robots has focused on either fast instruction-based methods which accept highly constrained input or gradual cross-situational learning methods, but not a mixture of both. In this paper, we present an integrated robotic system which allows for a combination of these methods to contribute to the task of learning the labels of objects in AI agents. We demonstrate the expanded word learning capabilities in the outcome system and how learning from both human-human and human-robot dialogues can be achieved in one integrated system.**

I. INTRODUCTION

Learning the mappings between the linguistic symbols and objects in real world is an instance of the problem of *grounded language learning* [1]. Past work on grounded language learning has mostly focused on the use of human-robot dialogues [2]–[7]. Teaching robots through structured human-robot dialogues (instruction-based methods) provides a fast and accurate approach for teaching a few word-object associations but will require many hours of training for teaching large sets of word-object associations.

Instruction-based word learning [4], [5], [7] usually capitalizes on identification of utterances whose syntax and semantics conform to a number of pre-specified structured representations, each representing a particular *definition instruction* for which the communicative goal of the speaker is also pre-specified. A definition instruction is used to map a novel word to an object, one of its parts or properties, an action, or any other concept which is pre-specified by the instruction as the communicative goal of the speaker. Instruction-based word learning enables fast and accurate word-object association learning while relying on a great amount of pre-specified details under constrained perceptual and linguistic circumstances.

Cross-situational approaches to word learning [8]–[13] on the other hand, accept naturalistic perceptual and linguistic input (⟨utterance,scene⟩ pairs) but they are usually slow due to capitalizing on gradual aggregation of cross-situational information across contexts in order to guide the process of mapping the words to their referents. Overall, cross-situational word learning offers a slow but flexible approach towards the acquisition of word-object associations.

Integration of these two types of word-object association learning methods into a single Cognitive Robotic Architecture (CRA), will enable the robot to retain the strengths of both approaches while making up for the limitations of each approach drawing on the information provided by the other approach. In this paper, we propose such a system which integrates both cross-situational word learning and instruction-based methods within the distributed integrated affect, reflection, cognition architecture (DIARC) CRA [14].

The rest of this work begins with the detailed description of the two word learning methods as well as the architectural modifications required to integrate them within the DIARC CRA. Then we discuss how the integrated word learning methods bootstrap the performance of each other and benefit from extra sources of information available in the CRA. Finally we close by discussing and demonstrating the new word learning capabilities achieved through the integration of the two methods.

II. INSTRUCTION-BASED WORD LEARNING (IBWL)

We use the DIARC CRA, and the methods described in [7] to implement instruction-based word learning. This section provides a high-level overview of that approach, and section IV describes how it has been extended to concurrently enable cross-situational word learning methods described in III. IBWL is initiated when the robot hears an utterance that contains a novel word as part of an utterance whose syntax and semantics match one of a set of predefined definition structures. Such a predefined definition relates the novel word to either an explicitly observable part of the environment, or some other concept which the robot already knows (e.g., “this object is a knife”, or “the silver part of a knife is the blade”). After a definition has been identified, it is used to create a symbolic semantic representation of the novel word. This semantic representation associated with linguistic information about the word inferred from the agent’s natural language understanding system, and a visual representation of the object, either generated from a snapshot of the environment, or through the composition of existing visual knowledge. The resulting system is able to learn all of the necessary information to understand a new object in a single exposure, when the object is either physically co-present with the human and robot, or describable in terms of concepts that the agent already understands. Objects learned in such a fashion have a persistent representation in the robot’s knowledge base, and due to their symbolic nature, they can be easily shared with other robots with compatible knowledge representation frameworks.

While this type of learning is fast and comprehensive for a single object, it is not feasible to teach a robot a large variety of objects through such methods. Such teaching is expensive, it requires the human’s attention to be directed at the robot, and the robots attention to be directed at the human and the

[1] HRI Laboratory, Computer Science, Tufts University, Medford, MA 02111, USA

object that is being learned. The object must be describable using words and concepts that are understandable to the robot, in the absence of any distracting object physically co present. As a result only a single new word-object association can be learned at a time. Sharing knowledge between robots can speed up the process to some extent but fails to fully address the problem as each word-object association first has to be learned (strictly via spoken instructions in case of low resource languages) before it can be shared.

III. Cross-situational Word Learning (CSWL)

Cross-situational word learning is performed by integration of the incremental and memory-limited cross-situational word learning model proposed in [13], within the DIARC CRA. The rest of this section provides a high level overview of CSWL model used in our system. Please refer to [13] for more details. Section IV-B describes how the CSWL model operates in the DIARC CRA and interacts with the other components.

A. Speaker's Model of Utterance Generation

The input to cross-situational word learning are word learning *situations* (or *trials*) each of which consist of a pair of an utterance and a scene, where the utterance is an unordered set of words and the scene is a list of objects present in the scene. The learner assumes that in each situation, the speaker follows the generative process illustrated in Fig. 1 to produce an utterance (W_s) corresponding to the current scene (O_s) and using a lexicon (L). *Lexicon* refers to a many to many mapping between objects and their labels.

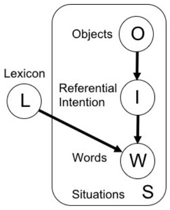

Fig. 1: The graphical model describing the generation of an utterance (W_s) in correspondence to the intention (I_s) and lexicon (L), and the generation of the intention (I_s) from the objects present in the scene (O_s). The plate notation indicates multiple copies of the model for different situations.

In each situation, the speaker is assumed to sample a subset from the power-set of all the objects present in the situation (O_s) as the *referential intention(s)* of the speaker (I_s). For each object in I_s, one word mapping is uniformly drawn from the lexicon and is added to the utterance. Words that are generated following this generative story and in reference to an object in the scene are characterized as *referential words*. The utterance may contain several non-referential words as well. Each word in the utterance is assumed to be referential with probability γ and non-referential with probability $1-\gamma$. Each referential word may be used non-referentially (in the absence of its referent in the scene) with $P_{NR} = \kappa < 1$.

The goal of the learner is to reverse this generative process and infer the lexicon used by the speaker. In doing so, the learner needs to find the $argmax_L P(L|C)$ according to the Bayes equation and the probability distribution that it defines over unobserved lexica (L) and the corpus of situations (C).

$$P(L|C) \propto P(C|L)P(L) \tag{1}$$

The model uses $P(L) \propto e^{-\alpha \cdot |L|}$ as a soft mutual exclusivity constraint to produce a preference for one-to-one mappings in the inferred lexicon. Marginalizing over all possible intentions in each situation we can rewrite the likelihood term $P(C|L)$ as:

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s|I_s, L)P(I_s|O_s) \tag{2}$$

Assuming that $P(I_s|O_s) \propto 1$ and that the words of the utterance are generated independently, we can rewrite the term $P(W_s|I_s, L)$ as:

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma)P_{NR}(w|L)] \tag{3}$$

B. The Incremental Learning Algorithm

The learning algorithm is composed of two major components: (1) inferring the context-appropriate parts of the speaker's lexicon (a mini-lexicon) in each situation, and (2) integrating the new mini-lexicon in the previous lexicon, while performing conflict resolution over alternative mappings. Inferring the maximum a posteriori (MAP) mini-lexicon, subsequently has two components: generating mini-lexicon proposals (groups of word-object mappings) and scoring the generated mini-lexica. Scoring is performed by computing the un-normalized posterior probability of the mini-lexicon proposals based on Eq. 1. We refer the reader to [13] for more details about the learning algorithm.

IV. CRA Integration

The configuration of DIARC used in this work (Figure 2) is based on the architecture described in [7]. A new Word Learning (WL) component has been added which implements the described CSWL. With this new component come new connections between components, and new types of exchanged messages across those connections. Information from the outside world enters the system through the Automatic Speech Recognition Component (ASR), and Vision Component (VIS). ASR converts spoken utterances into text which are passed to the Natural Language Understanding Component (NLU). NLU performs semantic and syntactic parsing on these utterances. NLU sends the utterance text, with Part of Speech (POS) tags, to WL, and the semantic representation of the utterance the Dialogue Management Component (DM). For a given utterance, the message to WL begins the process of CSWL, and depending on its contents, the message sent to DM may begin the instruction based word learning process.

The CSWL process involves ASR, NLU, WL, VIS, and the Belief State component (BEL). The instruction based word learning process involves ASR, NLU, DM, BEL, VIS, and the Goal Management Component (GM). The demonstrations in Sections V-A and V-B which involve IBWL, CSWL as well as action execution involve all of the components in Figure 2. The following sections describe the specific parts of the architecture which have been updated to enable the integration of IBWL and CSWL in the DIARC CRA.

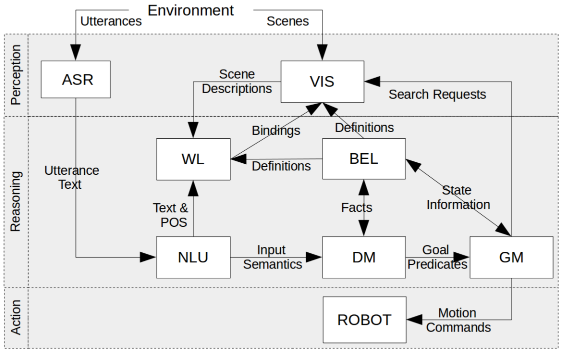

Fig. 2: The DIARC components used in this configuration, and the messages sent between them.

A. Natural Language Understanding

When a DIARC agent hears an utterance, whether it was spoken directly to the robot, or overheard from the environment, it is translated from an acoustic signal to text by the ASR component. This utterance text is then sent to the NLU component, which converts it into a semantic representation which is used by the rest of the system. It does this in three steps. First the syntactic structure of the utterance is determined using a Combinatory Categorical Grammar (CCG). This process determines the syntactic type of every word in the utterance, and how it relates to the rest of the utterance. When an utterance contains one, or a few, unknown words it may be possible to infer their syntactic type based on the grammar of the rest of the utterance, and an assumption that the utterance that the robot has heard is grammatical [7]. When the syntactic parsing step is completed, the words in the utterance and their POS tags are sent to the WL component.

After the syntactic structure of the utterance is determined, its semantic representation is derived from a set of lambda calculus rules which correspond the CCG rules used in the syntactic parsing [7]. No semantic representation will be assigned to novel words at this point. NLU generates a unique token for each novel word. The semantic representation of these novel words would be identified by other components (WL, or GM and VIS). Table. I provides example syntactic and semantic representations assigned to known words used in our demonstrations.

The final step of processing in NLU is pragmatic inference which converts the raw semantics of the utterance to a semantic representation of the speaker's intent. These intent semantics are sent to DM which determines how they should be handled, asserted to BEL in the case of facts and statements, and submitted to GM as goals in the case of commands or questions. It is worth noting that WL is updated with the utterance text and POS information before DM receives the utterance semantics. So in cases where the robot hears utterances that do not directly obligate it to act (for example overhearing two humans discuss a task), it is able to update its knowledge about the meanings of the words in the utterances without acting inappropriately.

Label	Syntax	Semantics
robot	N	$robot$
the	NP/N	$\lambda x.x$
object	N	$object$
is	(S/NP[a])\ NP	$\lambda x \lambda y.instanceOf(x,y)$
a	NP[a]/NP	$\lambda x.x$
an	NP[a]/NP	$\lambda x.x$
point to	C/NP[PP]	$\lambda x.pickup(?ACTOR,x)$
I	N	$?INTERACTOR$
will	(C\ N)/(C\ N)	$\lambda x.will(x)$
next to	(NP/NP)\ NP	$\lambda x \lambda y.nextTo(x,y)$

TABLE I: Parse rules used in demonstrations. S refers to a statement such as *S(human,robot,instanceOf(object,plate))* and C refers to a command such as *C(human, robot, pointTo(robot,plate))*.

B. Word Learning

WL manages the agent's knowledge of word-object mappings which are learned via CSWL. WL receives the utterance text with POS tags from NLU. Every time WL receives an utterance from NLU, it queries VIS for the list of visible objects in the scene. Then WL uses POS tags and the word-object associations learned via IBWL to trim its ⟨utterance,scene⟩ input received from NLU and VIS. WL uses the POS tags to determine the non-referential words (words whose POS is known and is not N) of the utterance and remove them from the input utterance. In addition to that, WL removes the word-object associations learned via IBWL from its input utterance and scene. Then WL utilizes the trimmed ⟨utterance,scene⟩ input to update its lexicon which will be shared with VIS upon the completion of the update process. The lexicon maintained by WL represents the agent's current implicit understanding of the word-object associations, while the word-object associations learned via IBWL represent explicit knowledge acquired by the robot and are treated as ground truth which should be protected from WL updates. That said, both types of acquired mappings can be immediately used when the agent is obligated to act (see Section V-A and Section V-B), while the mappings stored in WL lexicon are subject to modification in future encounters.

C. Belief Modeling

When the robot learns a new word-object association through IBWL, it stores the definition of that object in BEL and considers the learned association as ground truth which should be protected against modifications by CSWL (updates to the lexicon maintained by WL). In doing so, every time a new object definition is stored in BEL, BEL notifies WL which maintains a list of such objects and their word associations in order to to remove those words and objects from the input ⟨utterance,scene⟩ pairs it receives. Therefore, the word-object associations learned via IBWL are used towards reducing the number of words and objects input to CSWL which in turn serves to improve the word learning results from CSWL due to decrease in the referential ambiguity of its inputs (See Section V-C).

D. Vision

The DIARC vision component (VIS) is modular framework responsible for object detection and tracking. In the configuration presented here, VIS uses an Xtion sensor to process RGB-D data and is capable of detecting objects from a diverse set of object categories using a variety of image processing and detection modules. Beyond detection and tracking, another critical aspect of VIS is the ability to advertise its capabilities to the rest of DIARC. This advertisement is done through simple quantifier-free first-order predicate representations, specifying what kinds of object categories and properties VIS is capable of detecting (e.g., red(X), mug(X), on(X,Y)). In order for a DIARC component to make use of VIS capabilities, a request must be made in the form a quantifier-free first-order predicate representation. VIS then takes this request and attempts to start a visual search (if one is not already running) satisfying the entire request. Requesting components can then retrieve search results in the form of partial scene graphs which contain meta-data about object categories, object parts, and object properties.

Besides object detection and tracking, VIS is also implicated in both cross-situational and instruction-based learning. During instruction-based learning, VIS is notified by BEL of new definitions (e.g., the object is a plate) in predicate form (e.g. instanceOf(object(X), plate)). From this definition, VIS attempts to start a visual search for the referent object (i.e., object(X)) and find the corresponding object in the scene satisfying the request. If found, this RGB-D data of the segmented object is used to dynamically build a model of the object, attaching the appropriate label (e.g., plate). In the case of cross-situational learning, WL makes a request to get the visual search results of all objects in the current scene (from all currently running searches) for each utterance. Currently, there is an assumption that VIS is able to detect and track all objects that are used during CSWL even though a specific category label might not be known. Once WL has updated its lexicon, it shares the new lexicon with VIS, which is used to update its internal mapping of predicates to object categories maintained by VIS.

V. Demonstrations: Expanded Word Learning Capabilities

Integration of cross-situational and instruction-based word learning methods in the DIARC CRA, produces a robotic system which can capitalize on a wide variety of observations as word learning input to expand its vocabulary. Furthermore, cross-situational learning can benefit from context-disambiguation capabilities which draw on the lexical information provided by the syntactic representation of unknown words as well as known word-object associations learned via instruction-based word learning.

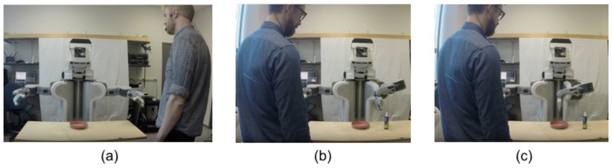

Fig. 3: (a) IBWL: "Robot, this object is a plate", (b) CSWL: "Robot, point to the bottle", and (c) CSWL: "Robot, point to the plate". The robot knows the meaning of "point to". Link to the demo: `https://bit.ly/2NtR1pf`.

We demonstrate the expanded word learning capabilities in two robot proof-of-concept demos and a simulation experiment which highlights the contribution of the context disambiguation capabilities available in the outcome integrated system in improving the cross-situational word learning results. The first demo is a proof-of-concept demonstration which highlights a specific learning capability (*zero-shot word learning*) which would have not be possible given an entirely instruction-based approach or entirely cross-situational approach. The second demo is a proof-of-concept demonstration which showcases the ability of the robot in capitalizing on IBWL to learn from its direct interactions with humans as well as capitalizing on CSWL to learn from human-human dialogues without direct engagement in human conversations.

A. Zero-Shot Word-Object Association Learning

The first demonstration illustrates zero-shot learning as a new capability within the DIARC CRA. This capability is resulted by the refinement of the CSWL inputs capitalizing on the syntactic information associated with unknown words as well as the knowledge of word-object associations learned via IBWL. The demo starts with an IBWL trial to explicitly teach the robot the label of the object PLATE, followed by a CSWL trial where a novel label is uttered in a known command ("point to") at the presence of a novel object in the scene. Fig. 3 illustrates the three trials of the demo and includes a link to its video.

The first interaction starts with the word "robot" in the utterance which triggers IBWL. Then "this object is a plate" is parsed into a form representing IBWL semantics, which is asserted to BEL and leads to mapping "plate" to the object PLATE via IBWL. This IBWL definition is then communicated to both WL and VIS as described in Section IV. Communicating the association between the word "plate" and the object PLATE to CSWL leads to the removal of the word "plate" as well as the object category PLATE from the input utterance and scene to the CSWL. This leaves CSWL with an empty utterance and an empty scene.

Therefore CSWL will not learn any new mapping but the robot learns that the word "plate" refers to the newly learned object category PLATE through IBWL. In the next trial, the object BOTTLE is added to the scene and the robot hears "point to the bottle". This utterance does not conform to instruction-based learning semantics, so only CSWL is used in this situation. First the utterance and scene descriptions are refined by removing word-object associations learned via IBWL along with non-referential words (with POS tags other than N). This leaves CSWL with one novel word in the utterance "bottle" and one novel object (BOTTLE) in the scene. CSWL instantiates a mapping between "bottle" and BOTTLE and adds this mapping to its lexicon, which then will be shared with VIS. Here, the semantics of "point to the bottle", generated by NLU, represent a goal for the robot, which is submitted to the Goal Manager (GM) to be executed. During execution, GM queries VIS for the object of interest, which VIS correctly identifies because of the updated lexicon provided by CSWL. The robot then uses the location of the object of interest to successfully point to the bottle, demonstrating the ability to learn a new word-object association in zero-shot via CSWL. Note that, only word-object associations learned via IBWL (explicit instruction) are removed from the input for CSWL. The word-object associations learned via CSWL (in the CSWL lexicon) are not removed from the input but the soft mutual exclusivity constraints devised in the CSWL model produce a preference for one-to-one mappings in the CSWL lexicon which discourages the model from mapping a novel word to a previously labeled object. The demo concludes by instructing the robot to "point to the plate", demonstrating that both objects have been learned correctly, one through instruction-based and one through cross-situational learning.

B. Continuous Acquisition of Word-Object Associations from Human-Human Dialogues

The second demonstration shows how the robot can learn from its direct conversations as well as overhearing the conversation of other agents present in the same physical context. A video of this demo is located here: `https://bit.ly/2IEWASV`. The robot starts with no knowledge of the label for the objects on the table. The demo starts with IBWL where the robot is directly addressed (via uttering "robot") and taught explicitly the definition of a plate. Note that During IBWL, there is no distracting object in the view point of the robot. IBWL is followed by CSWL in a series of human-human interactions observed by the robot.

The robot's vocabulary is sufficient to understand that the humans are talking about themselves ("I"), and to determine which parts of the utterance contain potential object labels (e.g., "the X"). Furthermore, it is able to infer (through NLU parse rules, see Table. I) the POS tags of the verbs and prepositions that are used ("take", "put", "near"), but it has no notion of what they mean. CSWL draws on the systems' knowledge of word-object associations learned via IBWL (e.g., plate-PLATE association), as well as the POS tags assigned to individual words to narrow down the potential novel labels and novel objects that correspond to one another. More specifically, CSWL removes the words learned via IBWL along with their associated referent objects from the input $\langle$utterance,scene$\rangle$. In addition to that, only words with the POS tag of N are preserved in the utterance and considered as potential labels for one of the objects present in the scene. This way, the input $\langle$utterance,scene$\rangle$ received by the CSWL is trimmed and its referential ambiguity is significantly reduced. Reducing the referential ambiguity of input serves as a context disambiguation approach and in Section V-C, we examine its effect on cross-situational word learning results.

At the end of the demo, the robot successfully points to three of the four objects ("box","mug", and "plate"), but fails to learn a correct mapping for "bottle". Such failures can be corrected by gaining more exposure to the word "bottle" in other contexts and diversifying the context in which the word "bottle" is used. [13] examined the effect of input order, referential ambiguity and exposure to more cross-situational data on word learning results in large simulation experiments.

Human1: Robot, this object is a plate.
Robot: OK.
Human2: I will take the bottle.
Human1: I will take the box.
Human2: I put the mug next to the plate.
Human1: I put the box next to the mug.
Human2: I put the bottle next to the box.
Human1: Robot, point to the bottle.
Robot: OK.
Human2: Robot, point to the box.
Robot: OK.
Human1: Robot, point to the mug.
Robot: OK.
Human2: Robot, point to the plate.
Robot: OK.

C. Context Disambiguation Capabilities

One of the new capabilities achieved in the current integration is the ability to disambiguate the context during CSWL, drawing on syntactic information provided by NLU and the information acquired by IBWL. The goal of this simulation experiment is to highlight the contribution of such disambiguation capabilities which reduce the referential ambiguity of the input for CSWL and improve its results.

To evaluate the effect of context ambiguity on CSWL results, we systematically varied (1) the number of distracting objects per scene and (2) the familiarity of distracting objects. Results are illustrated in Fig. 4. We used the probabilistic generative process in [15] for random generation of the utterances of a series of synthetic datasets (each with 100 trials). Then for each dataset, we manually generated the scene representations corresponding to the utterances of the dataset, where for each utterance the corresponding scene was composed of a list of the objects mentioned in the utterance. Then, we varied the ambiguity of the scene representations on two dimensions (1) the total number of unknown distracting objects encountered in the dataset and (2) the number of unknown distracting objects per scene. Specifically, we used 3 sets of distracting objects with 10, 30,

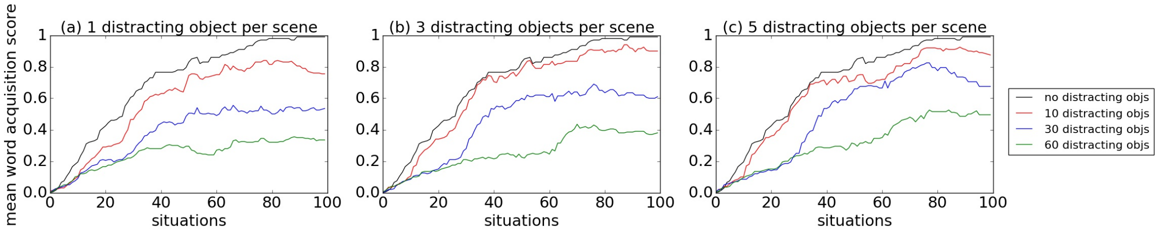

Fig. 4: Incremental word acquisition score on a given dataset varying (1) the number of unknown distracting objects per scene, and (2) the total number of unique unknown distracting objects in the dataset. The results are averaged over 10 runs.

60 new objects, none of which were used in data generation process (not mentioned in any of the utterances of the dataset). Using each set of distracting objects, we generated four variations of the original dataset correspondingly with 0, 1, 3, and 5 distracting objects added per scene. The frequency (which quantifies the familiarity) of a distracting object in a given dataset of size s, with a total of n unique unknown distracting objects and m distracting objects per scene is equal to $\frac{s}{(n/m)}$. For example, the frequency of a distracting object in a dataset, with 100 trials, total of 10 distracting objects and 2 distracting objects per scene is 20. we use *mean word acquisition score* [11] to evaluate the incremental performance during the simulation. G represents the gold standard lexicon and L represents the output model lexicon or generally speaking the target lexicon which is being evaluated.

$$mean\ word\ acquisition\ score = \frac{\sum_{\langle w,o\rangle \in G} P(w|o,L)}{size(G)} \tag{4}$$

Fig. 4 illustrates the incremental performance of the model on one of our randomly generated datasets as the ambiguity of the scenes and familiarity of the distracting objects are systematically varied by changing: (1) the number of distracting objects per scene, and (2) the total number of distracting objects in the dataset. Note that mean acquisition score improves upon receiving more input in all conditions and it converges to 1 in the absence of any unknown distracting objects. As can be seen, fixing the number of distracting objects per scene, using larger number of distracting objects in the dataset leads to a smaller acquisition scores and a slower rate of acquisition. This is due to the decrease in the frequency of the distracting objects in the dataset which lowers the context familiarity. On the other hand, fixing the total number of distracting objects in the dataset, using larger number of distracting objects per scene yields better acquisition scores which is due to the increase in the frequency (familiarity) of each distracting object. Similar trends were observed using other randomly generated datasets.

VI. Conclusion

We presented an integrated robotic system which capitalizes on a combination of fast instruction-based methods and gradual cross-situational learning methods to continuously learn new word-object associations from human-robot and human-human dialogues. In addition to that, the new integrated system utilizes the previously un-used syntactic information provided by NLU within the DIARC CRA, in the process of word-object association learning, which mirrors syntactic bootstrapping in language acquisition literature.

We discussed and demonstrated the expanded word learning capabilities in the outcome system including: (1) zero-shot learning, (2) learning from human-human and human-robot dialogues, and (3) the added context disambiguation capabilities. The presented system, outperforms its entirely instruction-based counterpart [7] in its ability to learn from human-human dialogues without the direct engagement of robot in the conversation as well as the ability for zero-shot learning. We discussed zero-shot learning in contrast to one-shot learning. Zero-shot learning characterizes the kind of learning which allows for the immediate use of the acquired information in the absence of any training trial. We demonstrated how such capability can be achieved by CSWL relying on context disambiguation capabilities available in the current integration. Similarly, the presented system outperforms its entirely cross-situational counterpart [12] in its ability to learn from explicit instructions, as well as its context disambiguation capabilities drawing on the syntactic information provided by NLU as well as the explicit knowledge acquired via IBWL.

Instruction-based methods, despite being fast and providing highly reliable explicit information, rely on highly constrained inputs. On the other hand, cross-situational leaning methods, allow for learning from a wide variety of naturalistic input, but their acquired knowledge is subject to noise. The current integration treats the information acquired via IBWL as ground truth and protects it from further modification. Future work should focus on exploring ways by which implicit knowledge acquired by CSWL can be transformed into explicit ground truth knowledge. One such possibility might be the use of active learning methods to have the robot ask clarifying questions before internalizing the information acquired via CSWL as explicit ground truth.

Acknowledgment

This project was in part funded by Vienna Science and Technology Fund project ICT15-045.

References

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[2] S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, and A. Klein, "Training personal robots using natural language instruction," *IEEE Intelligent systems*, vol. 16, no. 5, pp. 38–45, 2001.

[3] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," *Def*, vol. 2, no. 6, p. 4, 2006.

[4] R. Cantrell, P. Schermerhorn, and M. Scheutz, "Learning actions from human-robot dialogues," in *RO-MAN, 2011 IEEE*. IEEE, 2011, pp. 125–130.

[5] C. Meriçli, S. D. Klee, J. Paparian, and M. Veloso, "An interactive approach for situated task teaching through verbal instructions," in *Intelligent Robotic Systems*, 2013, pp. 47–52.

[6] M. Al-Omari, P. Duckworth, D. C. Hogg, and A. G. Cohn, "Natural language acquisition and grounding for embodied robotic systems." in *AAAI*, 2017, pp. 4349–4356.

[7] M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt, "Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1378–1386.

[8] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, no. 1, pp. 39–91, 1996.

[9] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using speakers' referential intentions to model early cross-situational word learning," *Psychological Science*, vol. 20, pp. 578–585, 2009.

[10] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic computational model of cross-situational word learning," *Cognitive Science*, vol. 34, no. 6, pp. 1017–1063, 2010.

[11] A. Alishahi and A. Fazly, "Integrating syntactic knowledge into a model of cross-situational word learning," in *Proc. of CogSci*, vol. 10, 2010.

[12] S. Sadeghi, M. Scheutz, and E. Krause, "An embodied incremental bayesian model of cross-situational word learning," in *Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (icdl-epirob)*, 2017.

[13] S. Sadeghi and M. Scheutz, "Sensitivity to input order: Evaluation of an incremental and memory-limited bayesian cross-situational word learning model." in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2018.

[14] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the diarc architecture," in *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013, p. 66.

[15] S. Sadeghi and M. Scheutz, "Early syntactic bootstrapping in an incremental memory-limited word learner," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.