An Embodied Incremental Bayesian Model of Cross-Situational Word Learning

Sepideh Sadeghi
Department of Computer Science
Tufts University
Medford, MA 02155
Email: sepideh.sadeghi@tufts.edu

Matthias Scheutz
Department of Computer Science
Tufts University
Medford, MA 02155
Email: matthias.scheutz@tufts.edu

Evan Krause
Department of Computer Science
Tufts University
Medford, MA 02155
Email: evan.krause@tufts.edu

***Abstract*—Learning the meaning of words in noisy contexts with multiple unknown words in an utterance and multiple unknown objects in a scene is a typical part of language acquisition for infants. However, incremental word learning in ambiguous contexts is a challenging problem in artificial intelligence. While past models of cross-situational word learning benefit from full access to all learning situations and their statistical regularities to arrive at the right hypothesis, it is cognitively implausible for children to remember all word learning situations they encounter. Hence, we present an incremental Bayesian model of cross-situational word learning with limited access to past situations and demonstrate its superior performance compared to other baseline incremental models, especially under conditions of sensory noise in the speech and visual modalities. Then we embed our model in a cognitive robotic architecture and demonstrate the first robotic model capable of incremental cross-situational word learning.**

I. Introduction

[1] suggested that the process of word learning is guided by observing referents of words across different situations. Many computational models of *cross-situational word learning* have shown that cross-situational learning is a powerful and effective mechanism for learning the meaning of words. The rule-based model of [2] and more recent probabilistic models [3]–[7] all rely on the regularities of the co-occurrences of words and meaning elements and successfully learn word meanings from noisy and ambiguous data. The rule-based nature of [2] limits its adaptability to new or natural data as it is not possible to revise the meaning of a word once it is considered learned, which prevents the model from handling highly noisy or ambiguous data. [3] uses the original form of the automatic translation learning algorithm of [8], which, however, lacks cognitive plausibility as it is non-incremental and learns through an intensive batch processing of a whole training data. While these models successfully accounted for many behavioral phenomena, the correlations between words and objects is typically noisy as speakers often talk about a few of objects present in the scene, or objects that are not visible at the time of utterance. Therefore, word learning is frequently intertwined with the problem of understanding referential intentions of the speaker. [7] bootstrap the process of learning word meanings with the model's belief about the referential intentions of the speaker, modeling the problem of word learning as the joint acquisition of the speaker's referential intention and word meanings. They tested their model using annotated corpus data and demonstrated that it has competitive results in comparison with other models including the IBM Machine Translation Model I [8], the statistical machine translation model [3], and cross-situational word learning using co-occurrence frequency, conditional probability and point-wise mutual information. However, it is impossible to use their model in real-time learning systems, despite its functional performance and its success in accounting for the behavioral phenomena, because the model operates in "batch mode", ignoring the incremental nature of the input in real time systems. This also makes the model less cognitively plausible.

In this paper we first present an incremental version of the Bayesian cross-situational word learning model proposed by [7] and compare its functional performance with other incremental models under different noise conditions. We show that our model demonstrates superior performance and robustness to noise. Then we integrate our incremental word learning model in a cognitive robotic architecture (CRA) to demonstrate that a robot can learn words incrementally and adaptively from individual word learning situations. In each word learning situation, the human interactor utters a sentence in the presence of multiple objects (some of which may be distracting objects) in the point of view of the robot. We demonstrate how learning unfolds incrementally as the robots receives more word learning situations and also study the effect of noise in vision and speech recognition components in DIARC [9] (our CRA) on the word learning results.

II. Word Leaning Model

Our model reduces the problem of learning the meaning of words into the problem of learning the referents of words as oppose to learning a distributed semantic representation for each word. Furthermore, the model is limited to learning the referent of words with concrete object referents. The input to the model are word learning situations each of which consisting of a scene paired with an utterance, where the scene description consists of a list of objects present in the scene and the utterance corresponds to an un-ordered set of words (ignoring syntax). Both scene and utterance may as well be empty lists. The objects listed in the scene description are not necessarily the ones talked about by the speaker and

the model relies on what it already knows about the words and their object referents (*lexicon* which is a many to many mappings between words and objects) to narrow down the focus of attention and identify the referential intentions of the speaker. We use the term *referential intentions* in the rest of this paper to refer to the objects that are present in the scene and the speaker is talking about them.

A. Model Design

The model assumes that the speaker uses the generative story captured in Fig. 1 to produce the words of utterance in each word learning situation. The learner has to reverse this generative story to infer the lexicon used by the speaker. In doing so, the model has to find the MAP (maximum a posteriori) lexicon by marginalizing over all possible referential intentions in each word learning situation. Fig. 1 represents the word learning variables and their probabilistic dependencies.

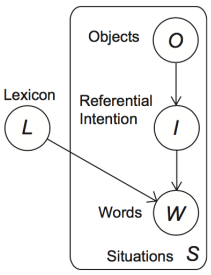

Fig. 1. The graphical model describing the generation of words (W_s) from the intention (I_s) and lexicon (L), and the generation of the intention (I_s) from the objects present in the scene (O_s), where s indexes the situation. The plate indicates multiple copies of the model for different situations (utterance-scene pairs). Image from [7].

In each situation, the model uniformly samples 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). Then for each object in I_s, one word is generated using the existing word-object mappings in lexicon L. Words generated in that way are referential words, but the utterance can include non-referential words (generated randomly) as well. Each word in the utterance is assumed to be used referentially with probability γ and non-referentially with probability $1-\gamma$. The non-referential use of words is quantified by P_{NR}, which is set to $\kappa < 1$ for words in the model lexicon (to penalize the non-referential use of words in the lexicon), and is set to 1 for other words. The referential use of a word w_i to refer to a particular object o_j is quantified by P_R which is the probability of w_i-to-o_j mapping as the outcome of a uniform draw from all the existing mappings for o_j in the lexicon. In each situation the model infers a mini lexicon corresponding to the entities (words and objects) in the current situation. In doing so, the model finds the MAP lexicon according to the Bayes equation and the probability distribution that it defines over unobserved mini-lexica (L) and the relevant corpus of situations including the current situation as well as the extracted ones from the lexicon (which share some entity with the current situation). The extracted situations are made of the existing (in the lexicon) mappings for each word and object in the current situation.

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

Given the probabilistic structure of the model and the fact that speaker's referential intentions are not observable, we marginalize over all possible intentions in each situation and 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:

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

We employ the equations above to find the MAP mini-lexicon in each situation to describe the relevant parts of word learning situations including the current situation as well as the relevant ones extracted from the lexicon. [7], on the other hand uses the equations above to find the MAP lexicon to describe all word learning situations in batch mode.

Another difference between our model and [7] is the function used to compute the prior probability of lexica. The goal for the prior is to serve as a mutual exclusivity constraint; however, the prior probability function used in [7] which takes the prior probability of a lexicon to be exponential in its size, generally favors small lexica over larger lexica even when the mutual exclusivity constraint is not violated. In order to make sure that we only discount lexica in which the mutual exclusivity constraint is violated, we take the prior probability of a lexicon to be exponential in the number of redundant links for the words of lexicon.

B. Incremental Learning Algorithm

We believe that only models with limited memory and computational resources qualify as scalable models which remain tractable as the amount of data grows. Following this assumption, we limit the model's memory of past observations to the word-object mappings stored in the lexicon and we define a set of constraints on incremental learning as follows. First, Incremental learning sees each situation only once (no iteration over data). Second, the model can only use the knowledge in its current lexicon and current observation for hypothesis generation and evaluation. Third, The model can only maintain a single global hypothesis across different situations motivated by recent findings in [10]. The model can make local revisions to this global hypothesis incrementally, as it receives more data. Our constraints on the data used for hypothesis generation and hypothesis evaluation may be too strict compared to the resources available to human

learners, but we believe that they are plausible approximation to the actual constraints that robot learners are subject to. Furthermore, in order to keep the Bayesian inference tractable, Bayesian inference is only applied locally for inferring the MAP mini-lexicon in each situation, where only context-appropriate word-object mappings available in memory are used for hypothesis generation and hypothesis evaluation.

Incremental learning has two components in our model: (1) inferring the MAP mini-lexicon in each situation, (2) merging the new mini-lexicon with the previous best lexicon found and resolving the potential conflicts. The process of inferring the MAP mini-lexicon, subsequently has two distinct components: (1) generating lexicon proposals, and (2) scoring the generated lexica. Scoring is performed by computing the relative posterior probability of the lexicon proposals based on Eq. 1. Generating lexicon proposals is guided by stochastic search techniques. The stochastic search in [7] is performed on all the possible links assuming full access to all word learning situations as inputs are processed in batch mode. Our stochastic search instead is performed only on the context-appropriate word-object mappings available in the memory (current lexicon and the current situation). Therefore our stochastic search is focused on small and relevant (to current situation) parts of the past observations. Focusing on smaller domains is in line with the "less-is-more" hypothesis [11] and, furthermore, more cognitively plausible.

Algorithm 1 Algorithm for updating the lexicon incrementally in light of a new situation.

procedure UPDATE(prevLex, situation, globalStats)
  $words \leftarrow unique(situation.words)$
  $objects \leftarrow unique(situation.objects)$
  $entities \leftarrow union(words, objects)$
  $links \leftarrow initLinks(words, objects)$
  $extractedLinks \leftarrow extractLinks(prevLex, entities)$
  $links \leftarrow union(links, extractedLinks)$
  $proposals \leftarrow initLex(nInit, links, globalStats)$
  $bestProposal \leftarrow bestScore(proposals, situation)$
  $situations \leftarrow union(situation, extractSit(prevLex, entities))$
  $PrevLexPart1 \leftarrow exclude(prevLex, entities)$
  $newMiniLex \leftarrow mutate(bestProposal, links, globalStats, situations)$
  $lexicon \leftarrow add(PrevLexPart1, newMiniLex)$
end procedure

Algorithm. 1 demonstrates the required steps for updating the lexicon learned by the model incrementally in light of new situations. *initLinks* initializes all possible word-object mappings using its input *words* and *objects*. *extractLinks* extracts the existing mappings of *entities* from the previous lexicon. *globalStats* contains some useful statistical measures such as point wise mutual information (PMI) of word-object pairs. These statistical measures are extracted from all situations observed so far and are incrementally updated as new situations are encountered. We use PMI of links as a *goodness heuristic* for links, employed in *initLex* and *mutate*. *proposals* is a list of *nInit* lexica, and each lexicon is a list of unique word-object pairs. *initLex* generates *nInit* new lexicon proposals in two steps: (1) sampling the length of the lexicon (we use a uniform distribution overall possible length values going from zero to the size of *links*), and (2) for each proposal, sampling $lexiconLen$ links from *links* according to a distribution created by normalizing exponentiated links' PMIs, where the exponent is the inverse of a temperature parameter. The temperature parameter can be used to adjust the stochasticity of the outcome of sampling, where higher temperature values make the outcome of sampling more stochastic. *bestScore* computes the posterior probability of its input lexica (hypotheses) given its input situation as data. Then, it samples one lexicon as the best one, according to a distribution created by normalizing the exponentiated un-normalized posterior probabilities for the input lexica, where the exponent is the inverse of another temperature parameter. *extractSit* extracts the existing mappings of items in *entities* stored in the previous lexicon (*prevLex*) and for each extracted mapping it creates a new situation with *link.word* as utterance and *link.object* as the scene description. The union of the current situation and the extracted situations are used as evidence when computing the posterior probability of mini-lexicon proposals. *exclude* removes all the existing mappings of items in *entities* from the previous lexicon, (except for those qualifying as the highest PMI mapping of a word available in memory), and stores the result in *PrevLexPart1*. *exclude* performs a strict mutual exclusivity constraint on the mappings suggested for each item in different situations. "mutation" of a lexicon refers to adding, deleting, or swapping a word-object pair to/from/in the lexicon. In each mutation step, the model generates 3 new mutated lexica from the base lexicon, using all three mutation moves. It, then uses each of the new generated lexica as the base lexicon for the next "mutation step". Therefore in two mutation steps we will have 3^2 lexica, which are the mutated versions of the base lexicon. *mutate* takes *bestProposal* as the base lexicon and applies the three mutation moves described earlier on it recuresively for *nStep*. After *nStep* mutations are completed, it evaluates the mutated lexica (3^{nStep} lexica) as well as the previous lexicon *bestProposal* using *situations* as data and selects one lexicon as the best one, by sampling a lexicon according to a distribution created by normalizing the exponentiated un-normalized posterior probabilities for the input lexica (the exponent is the inverse of the temperature parameter). *mutate* repeats these steps for *nIter* number of times and returns the result. Finally, *add* adds each word's mapping in *newMiniLex* to *PrevLexPart1* unless there is an alternative mapping for that word in *PrevLexPart1* with twice the PMI value of that mapping in *newMiniLex*.

C. Stochastic Search

The search for the best lexicon is partly guided by a heuristic search (*initLex* employs PMI of the links as the goodness heuristic), and partly by local optimization (mutating the lexica to maximize the posterior probability in *mutate*).

This optimization is local since it only tries to maximize the posterior probability given parts of all observations. The [7] model differs from our incremental model in that it combines heuristic based search with global optimization (maximizing the posterior probability of lexicon given all data). Global optimization is not a choice in the incremental model as past data is no longer available. However, the knowledge in model lexicon serves as model's interpretation of the past observations and the incremental model uses such knowledge as a proxy to past observations, by means of extracting situations from the previous best lexicon to evaluate the new proposals and their mutations in light of a new observation.

Our search for the best lexicon is stochastic due to employing a stochastic accepting criterion (sampling) for selection of the best lexicon proposal and links to be suggested to the lexicon. This way of choosing links and lexicon proposals introduces some stochasticity to the process, meaning that the model would not always select the links with the highest PMI value and lexica with the highest posterior probability in a greedy manner. We use two temperature parameters, one for lexicon acceptance and another one for link suggestion to modulate the degree of "greediness" in our model. The inverse of the temperature values is used as the exponent for the PMI value and the posterior probabilities when sampling a link or lexicon accordingly. Lower temperature values would magnify the score differences of links/lexica and makes the model's choices more greedy. On the other hand, higher temperature values smooths the score differences and makes the model choices more stochastic.

Unfortunately, the posterior distribution over lexica is highly irregular and finding the MAP lexicon requires smart ways of exploring the posterior distribution over lexica. Initializing new mini lexica in each situation, serves as global search in the space to perform semi-smart and yet random restarts. Mutations performed on the best of those mini-lexica serves as local search in the space around the best random start. Furthermore, our stochastic lexicon acceptance criterion along with our random restarts serve to avoid getting stuck in local maxima.

III. Evaluation

We integrated our model as a new component in the cognitive robotic architecture DIARC [9] to be able to perform two evaluations: (1) a systematically varied noise evaluation, and (2) an embodied human-robot interaction proof-of-concept demonstration.

A. Sensory Noise Evaluation

To evaluate the model under various amounts of sensory noise (i.e., visual object recognition and speech recognition), we ran a simple DIARC setup consisting of three components: (1) a simulated speech recognition component, (2) a simulated visual object detection component, and (3) a word learning component. During this evaluation, the ground truth of both the speech recognition and visual detection components was independently and systematically varied with pre-determined amounts of noise, i.e., for each visual object detected or word heard there was a *n%* chance that the word or object was misclassified. We considered five noise levels: 0%, 5%, 10%, 15%, and 20%. Table I and Table II show our evaluation data. We evaluate the word learning results by examining the precision, recall and F-score of the lexicon found by the model. To evaluate the incremental performance of the model, we use *average word acquisition score* ($P(object|word)$) [6] over all the words in the gold-standard lexicon as the performance measure. Fig. 2 demonstrates how incremental learning unfolds over time using *average word acquisition score* as the performance measure. As can be seen, average word acquisition score improves upon receiving more data and the learning curve converges to 1, slightly before receiving the 40th situation which indicates the stability of the word-object mappings learned by the model.

Most cross-situational models assume that infants are capable of correct object categorization and utterance segmentation prior to the word learning process, assuming that infants are capable of ignoring the individual differences between different instances of an object like DOG and regarding all instances as a DOG object, which along with the observed statistical regularities of the word *dog* and the object DOG, across different situations, allows for cross-situational word learning. This is yet another simplifying assumption as the process of object categorization probably interleaves with the word learning process. We examined the robustness of our model to noise in vision and speech recognition by systematically adding noise to the inputs from these two components and evaluating the mean F-score of the best lexicon found by the model, averaged over 10 runs. Fig. 3(a) demonstrates the behavior of our model under different noise conditions. The noise percentage value specifies the probability by which a word in the utterance or an object in the scene may be recognized incorrectly. Incorrect recognition of a token (word or object) in speech recognition and vision components refers to returning a(n) known/unknown token other than the actual one received by the sensory components.

For the sake of comparison, we implemented several incremental models of cross-situational word learning (association frequency (Eq. 4), conditional probability $P(object|word)$, conditional probability $P(word|object)$) mainly to provide a baseline expectation for the results produced by an incremental model. Our baseline incremental models are non-Bayesian statistical models of cross-situational word learning which similar to our model do not have full access to all data.

$$P(word, object) = \frac{Count(word, object)}{\sum_i \sum_j Count(word_i \cdot object_j)} \quad (4)$$

The best lexica found by the non-Bayesian models are a collection of word-object pairs with the highest heuristic (i.e, $P(object|word)$) score. We varied the number of links included in the best lexicon found by these models and reported the lexicon with the best F-score in zero noise condition. Note that the F-score reported for the incremental model in zero noise condition is not the best value, instead it is the average

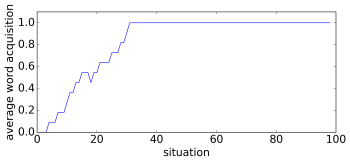

Fig. 2. Average word acquisition score (over time) for all the referential words in the dataset used for sensory noise evaluation (when noise=0). The summary statistics of the lexicon found by the model is precision:0.846, recall:1, F-score:0.916.

F-score value over 10 runs. For the experiments with non-Bayesian models under noise, we fixed the length of their lexicon at the length which gave the best F-score under no noise condition. Then we ran each model under different noise conditions for 10 times and used the averaged F-score values to generate the heatmaps.

Fig. 3(a) demonstrates the behavior of our model under noise. Fig. 3(b), Fig. 3(c), and Fig. 3(d) demonstrate the behavior of non-Bayesian models under noise. As can be seen, our model exhibits more robustness to noise compared to other non-Bayesian models, as the range of the mean F-scores values reported in Fig. 3(a) is smaller than that of other models. Furthermore, the least mean F-score value reported for our model (0.76) is much higher than that of other models (0.28, 0.55, 0.2) as can be seen in Fig. 3. Another drawback of the non-Bayesian models is that the performance of these models depend on the number of links allowed in their lexicon. In a real time system, where the dataset is not known in advance and is received incrementally over time, there is no way for setting the lexicon length to its best value.

Our model is more sensitive to noise in utterance (from speech recognition) compared to noise in object recognition (from vision). This asymmetry is due to the fact that in each situation, the model assumes that all referential intentions are equally likely. Therefore, when marginalizing over all referential intentions, the mis-recognition of an object can be smoothed out by referential intentions which exclude the mis-recognized object. On the other hand, the mis-recognition of a referential word with a non-referential word cannot be smoothed out as the probability by which a word can be used referentially (γ) is not equal to the probability by which it can be used non-referentially. We used $\gamma = 0.7$ in our simulations. Similarly, the model which uses conditional probability $P(object|word)$ as goodness heuristic for choosing word-object mappings, is more sensitive to noise in utterance, since utterance serves as independent variable in this model and noise in that can severely affect the performance. The model which uses conditional probability $P(word|object)$ as goodness heuristic for choosing word-object mappings, is more sensitive to noise in object recognition since scene serves as independent variable in this model.

B. Robot Proof-of-Concept Experiment

To demonstrate that our model is capable of learning incrementally in real-world contexts, we embedded the model in a subset of the cognitive robotic DIARC architecture [9] where we exchanged the simulated speech recognition and visual object detection components used in the sensory noise model evaluations with components capable of processing raw speech [12] and point cloud data. Additionally, we integrated a speech production component (allowing the robot to provide verbal feedback) and robot manipulation component (allowing the robot to point to target objects int the environment). Fig. 4 shows a high-level view of the DIARC configuration.

Our robot demo can be viewed by clicking on Fig. 5 or following the link in the figure caption. This demo illustrates how incremental word learning unfolds over time. The robot starts with an empty lexicon (no known word-object mappings). The human interactor then starts to teach new words through a series of word learning situations (utterance-scene pairs). The interactions between the human and the robot fall into two categories: (1) training and (2) testing. Any interaction in which the sentence uttered by the human starts with a word other than "point" is a training interaction (word learning situation). The robot updates its lexicon each time it receives a new word learning situation followed by uttering "OK". Any interaction in which the human interactor utters "point to the X" is a testing interaction which is used to examine the robot's knowledge of words (e.g., the word X). If the robot has at least one word-object mapping for the word X in its lexicon, it uniformly draws one of those mappings and points to the object in the drawn mapping while uttering "here it is". Otherwise, the robot responds "I don't know what that is". The word learning situations used in our demo include single word utterances (e.g., "knife") as well as complete sentences (e.g., "look at the knife"). The human interactor changes the scene by configuring the objects on the table.

TABLE I

THE MINIMUM, MEAN AND MAXIMUM NUMBER OF OBJECTS AND WORDS IN EACH WORD LEARNING SITUATION FROM THE DATASET USED FOR SENSORY NOISE EVALUATION. THE DATASET CONSISTS OF 99 SITUATIONS, WITH 33 UNIQUE SITUATIONS REPEATED THREE TIMES. THE REPETITION OF WORD LEARNING SITUATIONS IS INTENTIONAL TO EXAMINE AND DEMONSTRATE THE STABILITY OF THE MODEL AS IT RECEIVES MORE INPUTS.

	Min	Mean	Max
Number of Objects	2	2.45	4
Number of words	2	3.3	4

TABLE II

EXAMPLE DATAPOINTS FROM THE DATASET USED FOR SENSORY NOISE EVALUATION. EACH ROW REPRESENTS A SINGLE WORD LEARNING SITUATION. OBJECTS ARE ENCODED IN UPPER CASE LETTERS AND WORDS IN LOWER CASE LETTERS.

Utterance	Scene
bowl next to cup	BOWL,CUP,KNIFE
bowl next to cup	BOWL,CUP
look bowl	BOWL,KNIFE

IV. DISCUSSION AND CONCLUSION

We presented an incremental and adaptive word learning model integrated into our cognitive robotic DIARC architecture and demonstrated how the model on a robot can

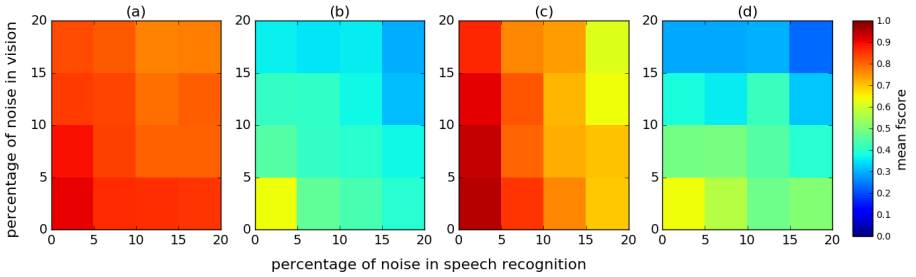

Fig. 3. The heatmap of mean F-score values (averaged over 10 runs) for the lexica found by (a) our proposed incremental model, (b) the association frequency model, (c) the conditional probability $P(object|word)$ model, and (d) the conditional probability $P(word|object)$ model, under different noise conditions.

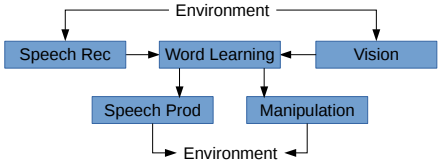

Fig. 4. High-level DIARC architecture for proof-of-concept demonstration.

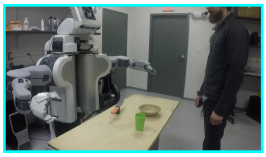

Fig. 5. Link to the demo: http://tiny.cc/68x5jy.

learn new words incrementally from word learning situations consisting of utterance-scene pairs. Our model departs from [7] mainly in its specification of information selection for hypothesis generation and hypothesis evaluation. The memory of our model is limited to the word-object mappings stored in the lexicon and the single situation is sees at each point in time. Furthermore, to keep the Bayesian inference tractable, our model is not fully Bayesian (only locally in the context of a single situation, where only context-appropriate word-object mappings available in memory are used for hypothesis generation and hypothesis evaluation). We compared the word learning results of our model with several non-Bayesian statistical models as baseline incremental models and showed that the model exhibits superior performance and robustness to noise in comparison with the baseline incremental models. The addition of noise in the utterance and scene descriptions injects noise in the correlations between objects and words which causes the baseline incremental models to fail. We believe that our model exhibits more robustness to noise due to two factors. First, it distinguishes between referential and non-referential words which allows the model to filter out mis-recognized words by excluding words from the lexicon that were used without a consistent referent. Second, the model allows for each and every object recognized by vision to be either present or absent in the referential intentions of the speaker. This helps to filter out the mis-recognized objects for which there is no reference in the utterance.

In our robot experiment, we used a limited number of objects (knife, cup, bowl) and words (look, at, the, cup, bowl, knife), to keep the demo short while demonstrating how incremental word learning unfolds over time. Further studies where the human interactor chooses how to naturally teach the words to the robot are required to examine the fitness of our model for real-time human-robot interactions. Furthermore, we plan to extend the current word learning model to learn action-verbs in addition to the nouns with concrete object referents.

V. Acknowledgments

This project was in part funded by Vienna Science and Technology Fund project ICT15-045 and by ONR grant N00014-14-0149.

References

[1] W. V. O. Quine, *Word and Object*. Cambridge, Massachusets: the MIT Press, 1960.
[2] 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.
[3] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing*, vol. 70, pp. 2149–2165, 2007.
[4] G. Kachergis, C. Yu, and R. M. Shiffrin, "An associative model of adaptive inference for learning word–referent mappings," *Psychonomic bulletin & review*, vol. 19, no. 2, pp. 317–324, 2012.
[5] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic incremental model of word learning in the presence of referential uncertainty," in *Proc. of CogSci*, vol. 30, no. 30, 2008.
[6] A. Alishahi and A. Fazly, "Integrating syntactic knowledge into a model of cross-situational word learning," in *Proc. of CogSci*, vol. 10, 2010.
[7] 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.
[8] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
[9] 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.
[10] T. N. Medina, J. Snedeker, J. C. Trueswell, and L. R. Gleitman, "How words can and cannot be learned by observation," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9014–9019, 2011.
[11] E. L. Newport, "Maturational constraints on language learning," *Cognitive science*, vol. 14, no. 1, pp. 11–28, 1990.
[12] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The cmu sphinx-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, vol. 1. Citeseer, 2003, pp. 2–5.