Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Bayesian Active Learning-Based Robot Tutor for Children's Word-Reading Skills

Goren Gordon and **Cynthia Breazeal**
Personal Robots Group, MIT Media Lab,
20 Ames Street E15-468
Cambridge, MA 02139
{ggordon,cynthiab}@media.mit.edu

Abstract

Effective tutoring requires personalization of the interaction to each student. Continuous and efficient assessment of the student's skills are a prerequisite for such personalization. We developed a Bayesian active-learning algorithm that continuously and efficiently assesses a child's word-reading skills and implemented it in a social robot. We then developed an integrated experimental paradigm in which a child plays a novel story-creation tablet game with the robot. The robot is portrayed as a younger peer who wishes to learn to read, framing the assessment of the child's word-reading skills as well as empowering the child. We show that our algorithm results in an accurate representation of the child's word-reading skills for a large age range, 4-8 year old children, and large initial reading skill range. We also show that employing child-specific assessment-based tutoring results in an age- and initial reading skill-independent learning, compared to random tutoring. Finally, our integrated system enables us to show that implementing the same learning algorithm on the robot's reading skills results in knowledge that is comparable to what the child thinks the robot has learned. The child's perception of the robot's knowledge is age-dependent and may facilitate an indirect assessment of the development of theory-of-mind.

Introduction

Socially assistive robotics is an emerging field which strives to create socially interactive robots that aid people in many different aspects such as care for the elderly and education (Tapus, Maja, and Scassellatti 2007; Fasola and Mataric 2013). Educational assistive robots interact with children and help them develop in an educational setting (Movellan et al. 2009; Saerbeck et al. 2010; Kory, Jeong, and Breazeal 2013; Fridin 2014). However, in order to tutor children in an effective manner, one must personalize the interaction and the curriculum to each child. This personalization can be achieved by proper assessment of the student's current skill and adaptation of the behavior and material taught to the student's level (Wise, Ring, and Olson 2000; Chambers et al. 2011). However, as opposed to standardized tests or tablet educational games, the robot should perform these tasks while being engaging (Brown, Kerwin, and Howard 2013), believable and social (Saerbeck et al. 2010), i.e. a companion for the child (Breazeal and Scassellati 1999). Thus, an integrated approach is required, that takes into account personalization, AI of the robot, curriculum and entertainment aspects.

In this contribution we have developed an integrated child-tablet-robot (Jeong et al. 2014) experimental design that attempts to address these specific concerns, in the context of learning how to read words. The setup consists of a social robot that is portrayed as a younger peer who wishes to learn to read, thus putting the child in the empowering position of the teacher. This also enables continuous assessment of the child, by having the robot ask questions as a less informed companion. Crucially, we developed an algorithm that allows the proper and efficient assessment of the child's word-reading skills and adapted the interaction between robot and child to each child's specific level. A unique tablet app was also developed to facilitate an engaging story-making game that can be co-played by the child and robot.

We show that the algorithm developed adequately represents the child's word-reading skills. Moreover, we show that across a wide age range (4-8 years old), the same algorithm properly assesses each child's skills. We further show that an active-learning approach, in which assessment is performed via maximizing expected information gain (MacKay 1992), results in a better assessment than other tested methods, e.g. random and border-of-knowledge assessments.

Finally, we ask whether children properly perceive their own teaching skills. This question is addressed by having the robot actually acquire knowledge with the same algorithm that assesses the child's knowledge, and testing the agreement between the child's perception of the robot's reading skills and its actual child-like learning. We show that this agreement, which relates to theory-of-mind (Astington and Jenkins 1999), is age dependent, wherein below 5 years of age, children wrongly over-estimate the robot's skill, and above that age, a more accurate representation of the robot's reading skill is developed.

Bayesian Active Learning Algorithm

The goal of the developed algorithm is to properly and efficiently assess a child's word-reading skills, which is here defined as the probability that the child knows how to read

words. We first describe the knowledge representation and update and then different assessment methods employed.

Knowledge representation and update

The knowledge is represented by a vocabulary of words, V, wherein each word $w \in V$ has an associated probability $p(w) = p(w^{\text{correct}})$, which represents the child's probability of correctly knowing how to read word w. Thus the algorithm should result in a proper matching between $p^{\text{algo}}(w)$ and the true knowledge of the child $p^{\text{true}}(w)$, $\forall w \in V$. Knowledge is updated according to Bayes's rule, by observing the child's reading of specific words and ascertaining their connectedness, i.e. $w_{\text{obs}}^{\text{correct}}, w_{\text{obs}}^{\text{wrong}}$ means the child reads word w_{obs} correctly and incorrectly, respectively. Thus the knowledge update, which occurs on all the words in the vocabulary, $w \in V$, is given by:

$$p(w|w_{\text{obs}}^{\text{correct}}) = p(w)\frac{p(w_{\text{obs}}^{\text{correct}}|w^{\text{correct}})}{p(w_{\text{obs}}^{\text{correct}})} \quad (1)$$

$$p(w|w_{\text{obs}}^{\text{wrong}}) = p(w)\frac{p(w_{\text{obs}}^{\text{wrong}}|w^{\text{correct}})}{1-p(w_{\text{obs}}^{\text{correct}})} \quad (2)$$

The vocabulary is constructed from all possible observed words during the interaction, such that $p(w)$, $\forall w \in V$ is assessed throughout the interaction.

The conditional probability $p(w_1|w_2^{\text{correct}})$ means, what is the probability the child knows how to read word w_1, given that she knows how to read word w_2. Alternatively, $p(w_1|w_2^{\text{wrong}})$ means, what is the probability the child knows how to read word w_1, given that she does not know how to read word w_2. We have constructed a heuristic for this conditional probability, based on a phonetic distance metric between the words, $d(w_1, w_2)$, defined as:

$$\hat{d}(w_1, w_2) = \sum_{i=1}^{|w_1|}\sum_{j=0}^{|w_1|-i} \frac{\delta(w_1[i:(i+j)] \in w_2)}{j+1} \quad (3)$$

$$n(w_1, w_2) = \sum_{i=1}^{|w_1|}\sum_{j=0}^{|w_1|-i} \frac{1}{j+1} \quad (4)$$

$$\delta(w_1[i:j] \in w_2) = \begin{cases} 1 & \text{letter sequence } i,...,j \\ & \text{in word } w_1 \\ & \text{appears in word } w_2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$d(w_1, w_2) = \hat{d}(w_1, w_2)/n(w_1, w_2) \quad (6)$$

where $|w|$ is the number of letters in word w, $\hat{d}(w_1, w_2)$ is the un-normalized distance metric and $n(w_1, w_2)$ is the normalization, which simply represents the total number of possible joint letter-sequences. This distance metric represents the fact that the more letter-sequences appear in both words, the more similar they are. Also, the effects of longer letter-sequences are smaller than shorter ones, which reflects that reading occurs from the letter to phonetic to whole words. Put differently, words with different letters are more dissimilar than words with dissimilar bi-letters.

This normalized distance metric is used in a heuristic exponential model to estimate the conditional probability functions:

$$p(w_1|w_2^{\text{correct}}) = \left(1 + e^{-d(w_1,w_2)/b}\right)/2 \quad (7)$$

$$p(w_1|w_2^{\text{wrong}}) = \left(e^{-d(w_2,w_1)/b}\right)/2 \quad (8)$$

The rationale behind these heuristics definitions is that the more similar the words, the more probable their informing one on another. Put differently, if a child does not know how to read a single letter, this will have a more dramatic effect on knowing how to read a word with that letter.

After each probability update, the probability of each word is bounded: $0.05 \leq p(w) \leq 0.95$, $\forall w \in V$. This is done in order to introduce some uncertainty in the final knowledge, such that even though many words were un/known, the un/certainty of reading other words is still bounded.

Assessment methods

During the interaction, the child is verbally "probed" to select a word out of several other presented ones (see Experimental Setup). For example, when a sentence appears, a single word is verbally spoken and the child is required to select it. We used four different methods to select which word, w, out of the list of words, W, to ask the child. The baseline method is a random selection of the word (Random), i.e. not using any acquired knowledge of the child's word-reading skill. The second method is a "border of knowledge" method (Border), wherein the word closest to 0.5 probability is selected, i.e. the word which the model estimates the child have a fifty-fifty chance of knowing and not knowing. This method challenges the child, but not too much. In addition, when the assessed child's knowledge is complete, i.e. $p(w) = 0.95$, $\forall w \in W$, the longest word is selected, so as to potentially challenge the child.

The third method is an "active learning" method (Active), which aims to optimally increase the knowledge of the child's reading skill. This method selects the word with the highest expected information gain (MacKay 1992), i.e. the word that maximally decreases the uncertainty in the probability distribution in the vocabulary. The measure is defined as follows:

$$I(w) = \sum_{w' \in V}\Big(p(w')D_{\text{KL}}(p(w'|w^{\text{correct}})||p(w')) + (1-p(w'))D_{\text{KL}}(p(w'|w^{\text{wrong}})||p(w'))\Big) \quad (9)$$

$$w_{\text{select}} = \text{argmax}_{w \in W} I(w) \quad (10)$$

where D_{KL} is the Kullback-Leibler divergence. The rationale behind this method is to select a word that reduces uncertainty weighted by the probability that she knows it or not. For example, a word that greatly informs as to the child's knowledge, but that is certain to be readable, is not likely to be selected.

The fourth method selects the word based on prior interaction (Select), e.g. assessing the child on a previously incorrectly identified word. This allows post-test assessment of the learning process.

Experimental Setup

Robotic platform

For the social robotic platform we used DragonBot (Setapen 2012), a squash-and-stretch Android smartphone based robot. The facial expression, sound generation and part of the logic is generated on the smartphone, which is mounted on the face of the robot. The robot appears to be a soft, furry, fanciful creature that is designed to engage children. DragonBot is a very expressive platform and has a large repertoire of possible facial expressions and actions. We installed a commercial child-like voice for the text-to-speech software on the smartphone, to facilitate a more generic and engaging interaction.

Initial assessment

During the introduction to the study, the child is asked to spell her name, and is informed that she is going to play word games with the experimenter and then with the robot. The first "word game" is the TOWRE word assessment test, in which the experimenter asks the child to read words from lists, as fast as she could, for 45 seconds. The raw TOWRE score is defined as the total number of correctly read words during these 45 seconds. We administered both sight and phonetic word lists, where the total raw score reported here is the sum of the two tests' raw score. We used this information, i.e. which words are read correctly and which are not, to initialize the knowledge of the child's reading skill. Thus, even prior to the interaction, there is some baseline of the child's reading vocabulary.

Robot interaction

The child sits next to a small table upon which there is a tablet and the robot, Fig. 1. The robot is "sleeping", i.e. its eyes are closed, and it is introduced as "Parle, a young robot that just learned how to speak and wants to learn to read". The robot awakens, yawns (an overt motion and sound), and introduces itself: "I am Parle, we are going to play word games together." The speech is interspersed with facial expressions and sounds, to create a more engaging interaction.

The first phase of interaction is a pre-test, during which the robot asks the child to teach it some words. It verbally asks the child to show it a word, e.g. "dragon", whereupon the word, and four distractors appear on the tablet. The child then needs to tap on the correct word. During this phase, the word is chosen according to the active-learning method, i.e. the word is the one that maximizes the expected information gain. The distractors are selected from the same vocabulary: two words which are most similar, i.e. smallest distance metric from the selected word; one word that the child should know, according to the assessment; and one word that the child should not know how to read. This is repeated ten times, to get a thorough assessment of the child's reading knowledge. The robot physically expresses excitement after each word taught and at the end of the assessment phase.

The second, and main phase of the interaction, the story phase, is based on an in-house developed app game that enables the child to co-create a story with the robot. The game contains three scenes, and several characters. The child can

Figure 1: The experimental setup.

move any character that appears on the screen. After each movement, a sentence is automatically generated using a novel auto-generation mechanism, which (i) randomly selects an adjective for the character; (ii) detects the closest other character for the story interaction; (iii) follows an XML script of the plot of the story; and (iv) uses an open-source natural language generation library to construct a full sentence.

The XML plot files are constructed in a generic fashion, such that (i) each character has a list of possible adjectives (e.g. red, big), motions (e.g. fly, jump) and speech (e.g. roar, squeak); (ii) the plot line is constructed of a sequence of movements, speech, feelings and resolutions; and (iii) the story conversation is constructed such that any sequence of character selection generates a coherent story line. The result of each movement is thus a full sentence that describes the progression of the story plot. After several such sentences, the scene changes and new characters are introduced, while some of the old ones are taken away. There are three scenes to the story, which ends with a final resolution sentence.

In the child-tablet-robot interaction, when the child moves a character, the robot first speaks the generated sentence, and then the sentence appears on the tablet above the scene. In 50% of the sentences, the robot expresses a shy face and asks the child to show it a word, e.g. "I don't know how to read the word dragon. Can you show it to me?". This resulted in an average of 11 words per interaction. The child is then required to tap on the correct word. Each tapping on a word on the tablet results in the tablet speaking that word. In this sense, the tablet is an informant, whereas the child and robot are both the students. If the child is correct, the robot becomes excited, i.e. moves in an excited manner, thanks the child and the story continues. If the child is wrong, the robot expresses frustration and asks the the word again. If the child is wrong again, the tablet shows the correct word in an emphasized manner and speaks it. The game then continues until the end of the story.

In the last stage of the interaction, the post-test, the robot again asks the child to teach it some words, similar to the pre-test phase. During this phase, the words that were asked

during the story phase are asked again, with priority to those the child got wrong, then those that she got right and finally the most informative words, i.e. the ones that maximize the expected information gain. A total of ten words are asked during this phase.

In order to increase believability and engagement with the robot, we inserted randomness to the expressions and sentences the robot asked, so as to avoid boring repetition. For example, during the pre- and post-test phases, the robot asked: "Can you show me the word X?", "That is a new word, X. Can you tap on it?", "I don't know the word X. Can you show it to me?" This increased diversity and randomness in the robot's behavior proved to be essential for the children's engagement, a major factor in educational interactions.

Conditions and subjects

There were two conditions in the experiment, which differed only in the story phase. In the "border" (B) condition, the border-of-knowledge method selected the words. Furthermore, we implemented the ceiling, i.e. if the child was assessed to know all the words, the longest one was selected. In the "random" (R) condition, the robot asked about a random word in the sentence.

More specifically, the methods employed in each conditions were the following: in the "border" condition, initial assessment employed "active-learning", story-phase employed "border of knowledge" and final assessment employed "select" method; in the "random" condition, initial assessment employed "active-learning", story-phase employed "random" and final assessment employed "select" method.

There were 49 children subjects of ages 4-8 of both genders. They were recruited via e-mail lists of family groups from the general surrounding. All participant's parents signed consent forms. Out of these subjects, only 34 completed the task with usable data ($n_B = 20, n_R = 14$).

Results

We first describe the general performance of the algorithm in representing each child's reading skill. We then analyze the results across ages and initial reading skills, as assessed by the TOWRE tests, followed by results pertaining to the different assessment methods. The child's actual learning during the interaction is described next, followed by discussion and implications for children's theory-of-mind development.

Algorithm represents child's knowledge

Each subject was asked 10 words during the pre-test phase, W^{pre}, 10 words during the post-test phase, W^{post} and a variable number of words during the story phase, ($|W^{\mathrm{story}}| = 11 \pm 1$ SE words), due to the probabilistic nature of the interaction. In order to assess the algorithm, we calculated the actual probability, p_i^{true}, that the child was correct over all asked words, $w \in W^{\mathrm{total}} = W^{\mathrm{pre}} \cup W^{\mathrm{story}} \cup W^{\mathrm{post}}$, i.e. during the entire interaction, and compared it to the average assessed word probability $p^{\mathrm{algo}}(w), w \in W^{\mathrm{total}}$:

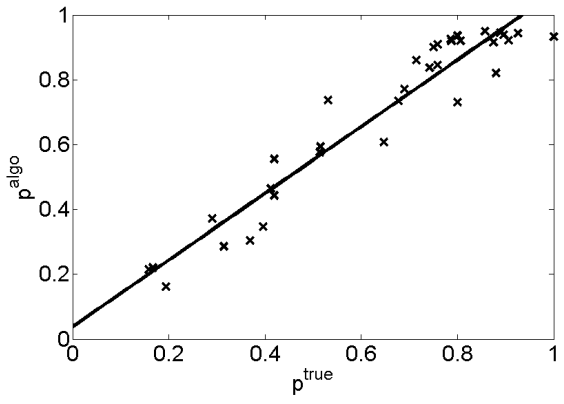

Figure 2: Algorithm assessed knowledge represented by child's probability of being able to read words, p^{algo}, as a function of child's actual word-reading skill represented by the probability to correctly identify a prompted written word, p^{true}.

$$p_i^{\mathrm{true}} = \frac{|W^{\mathrm{correct}}|}{|W^{\mathrm{total}}|} \quad (11)$$

$$p_i^{\mathrm{algo}} = \frac{1}{|W^{\mathrm{total}}|} \sum_{w \in W^{\mathrm{total}}} p(w) \quad (12)$$

where i represents each subject. In Fig. 2 we show the results of this comparison. As can be seen, the algorithm captures quite accurately each child's word-reading skills, measured by the probability of its ability to detect a prompted written word ($n = 34, R = 0.959, p < 0.01$).

Children word learning

The child's only opportunity for learning is during the story phase, in which the tablet game serves as the informant. The robot is framed as a younger peer who wants to learn from the child, and if the child cannot teach it, the tablet can. Thus, subjects can learn new words only during the short (5-10 minutes) interaction with the story game. Surprisingly, learning, defined as an incorrectly identified word during the story phase followed by a correct identification in the post-test phase, does occur. Out of the 34 subjects, 20 learned at least one new word, to the total of 29 new words over all subjects. The words learned ranged from "are" and "ball" to "castle" and "enchanted".

To quantify the learning process, we calculated how many wrongly identified words in the story phase (44) were correctly identified during the post-phase (29) and how many were not. This means that 66% of the words were identified, much greater than 20% chance during the post-test.

Personalization

To further study the algorithm's ability to personalize to specific children, we analyzed the algorithm's assessment error and the words learned as a function of both age and TOWRE score, Fig. 3.

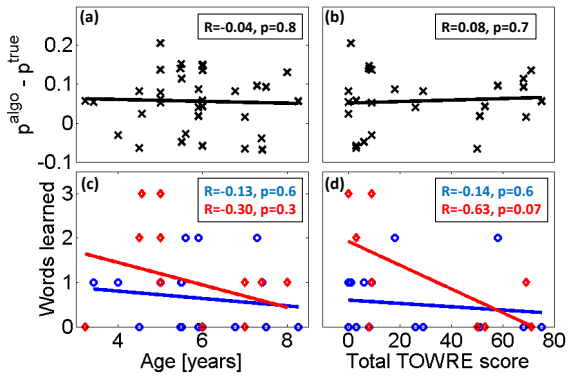

Figure 3: Personalization expressed as dependence of algorithm's assessment error (a,b) and words learned (c,d) on age (a,c) and TOWRE score (b,d). In (c,d), blue and red denote the Border and Random condition, respectively.

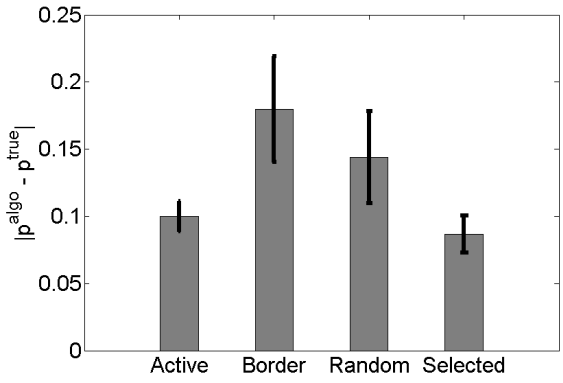

Figure 4: Average error between computed algorithm assessed word-reading skill and true word-reading skill for the different assessment methods (error bars represent SE). Active represents pre-test active-learning method; Border represents story border-of-knowledge method; Random represents story random selection and Select represents the post-test selected words.

The algorithm's error was calculated for each child as $p_i^{\text{algo}} - p_i^{\text{true}}$. Fig. 3(a) shows that the error is low and does not depend on age. In other words, the algorithm adapts equally well to every child, regardless of age. Furthermore, Fig. 3(b) shows that the error also does not depend on the raw TOWRE score, i.e. even though children differed greatly in their initial reading abilities, the algorithm correctly assessed their reading skill. It is important to note that the words in the TOWRE test were different than those used during the interaction, i.e. the algorithm transferred the information from the TOWRE test into the general vocabulary and then continued the assessment during the interaction. Furthermore, as can be seen, several subjects scored zero on the TOWRE test, i.e. they could not read a single word correctly from the lists. Still, during the interaction they were able to correctly identify some words, which were correctly assessed by our algorithm.

More evidence for successful personalization is the fact that the number of words learned in the Border condition does not depend on age or TOWRE score, Fig. 3(c,d), respectively. Without such personalization, i.e. in the Random condition, there is a much stronger dependence on both age and TOWRE score. While the absolute number of words learned is not significantly different between the two conditions, the Border condition, which employs personalization, is more uniform across subjects, thus guaranteeing that children learn regardless of their word-reading skills.

Assessment method analysis

During the interaction several assessment methods were employed, namely, Random, Border, Active and Select (see Assessment methods above). While all subjects were assessed with the Active (during the pre-test) and Select (during the post-test) methods, they were divided among the two experimental conditions (B and R).

During each assessment method applied, specific words were prompted and assessed. We calculated for each such method, for each subject (as noted above), the true probability of correctly reading a word, $p_i^{\text{true,method}}$ and the algorithm's prediction of it $p_i^{\text{algo,method}}$. We calculated for each method the error in algorithm prediction: $|p_i^{\text{true,method}} - p_i^{\text{algo,method}}|$. In Fig. 4 we show the mean and standard error of each method, averaged over subjects. While not significant, the Active method's error is small compared to the Random and Border ones, even though it is based on less knowledge. Put differently, during the story phase, the algorithm had more words to assess the child's knowledge, namely, those asked during the pre-test. Still, the assessment of the algorithm, due to the appropriate selection of words, outperforms the other methods. The Select method, which occurs in the post-test, is the best due to the words chosen, which were already asked before, and because it is at the end of the interaction.

Robot word learning and children's theory-of-mind development

We wanted to assess the children's perception of their teaching, i.e. how much children think they taught the robot. For this purpose, after the robot interaction we asked them whether they thought the robot could read six selected words. Put differently, after showing the robot words, did the children think the robot knew how to read? For each word the child could answer "Yes", "No" or "I do not know" if the robot knew how to read, for which we assigned $q_j = 1.0, 0.0$ and 0.5 probabilities, respectively, where $j = 1 \ldots, 6$. These represent $p_i^{child} = \frac{1}{6}\sum_{j=1}^{6} q_j$ for each subject.

Concurrently, we run the same Bayesian update algorithm on the robot. The initial vocabulary probabilities were set to 0.05, representing lack of knowledge. Then, according to Eq. (8), for each correctly taught word, w, the robot's knowledge was updated, whereas for each incorrectly identified word, the robot knowledge was updated twice, for the asked word and for the incorrectly identified one. The latter represented that the robot was misinformed on both words. For

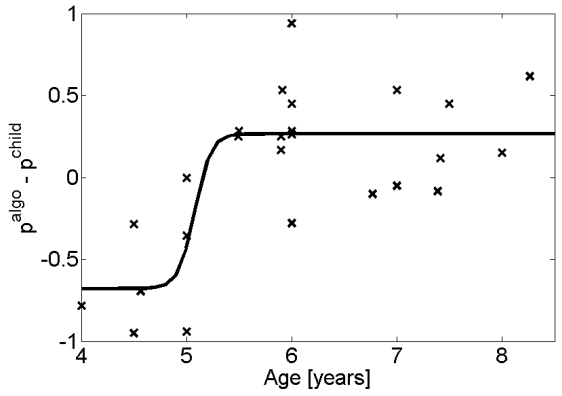

Figure 5: Error between children's perception of robot reading skills and actual reading skills as computed by the suggested algorithm, as a function of age.

each subject we then compared the average child's perception and the robot's learning according to our algorithm.

We hypothesize that the discrepancy between child and algorithm is age-dependent, as theory-of-mind develops around the age-range we studied. Fig. 5 indeed shows an age dependency, with a logistic fit suggesting a turning point around 5 years of age ($R^2 = 0.573$). The discrepancy plateaued around 0.2, comparable to the errors of the algorithm in the previous analysis. Finally, it shows that young children over-estimate the robot's learning capabilities.

Related Work

There is a large body of knowledge related to intelligent tutoring systems (ITS) (VanLehn 2011; Polson and Richardson 2013). A recent tutoring system, Guru, uses an interactive dialogue tutoring system that was modeled after 50 expert tutors in the domain of high-school biology (Olney et al. 2012). It was shown that the Guru system was comparable to human tutoring in post- and delayed-tests on the taught material, and significantly better than classroom-only teaching. In this contribution we employed a mathematical-based tutoring method, as opposed to human tutor-based, as well as targeting much younger students. Another study compared the effects of specific ITS employing an early literacy computer game as well as computer feedback-tutoring system on young children (Kegel and Bus 2012). It was shown that the literacy game was effective only when combined with the computer feedback tutoring. In this contribution we employed an active-learning tutoring, as opposed to only feedback, and introduced an embodied agent, as opposed to a computer-based tutor.

On the other hand, the development and research of robot tutors have recently been flourishing. RUBI-4 is a humanoid robot with articulated arms, an expressive face, and a tablet embedded in its midsection, with which it played simple vocabulary games with preschool children (Movellan et al. 2009). In this contribution, the robot was a peer companion that played an educational tablet game with children ages 4-8.

Studying the role of teaching as an effective learning strategy, Tanaka and Matsuzoe's robot played a verb-learning game, in which the experimenter asked either the preschool child or the robot to act out novel verbs (Tanaka and Matsuzoe 2012). They found that teaching the robot helped children remember the verbs, as well as inspiring further teaching-verbs play. We employed a similar empowerment concept, yet integrated a personalization algorithm to match each child's skill.

Recently, personalization of robot tutors, even via simple algorithms, has been shown to greatly increase tutoring effectiveness (Leyzberg, Spaulding, and Scassellati 2014). By contrast, our contribution used an active learning-based algorithm and focused on children's word-reading skills and not adult's game strategies.

Employing active-learning algorithms in a human-robot interaction paradigm has been shown to be more effective than supervised learning alone (Cakmak, Chao, and Thomaz 2010). Furthermore, subjects exhibited a more accurate performance estimate in the interactive modes using active learning than in the passive supervised learning mode. However, in this paper we have employed the active-learning paradigm on child's assessment and not interaction, as our main goal has been tutoring.

Conclusions

We have presented an integrated system that addresses the challenge of effective child-robot tutoring. Our system incorporates: (i) a social robot with an engaging social interaction, framed as a younger peer; (ii) an effective state-of-the-art assessment algorithm, based on a Bayesian active learning approach; (iii) an interactive story-making tablet app that allows an engaging co-play of robot and child.

By studying a large age-range of children subjects, we were able to show that the assessment algorithm and the entire system personalizes to different age-groups and initial reading skills. This short interaction resulted in actual word learning in an entertaining setup.

The results presented above indicate that another level of assessment is possible, i.e. the developmental level of the theory-of-mind of the child. The combination of an engaging, empowering and entertaining experimental paradigm facilitated the acquisition of the data. These results can only be obtained in an integrated system, with a social robot, an effective learning algorithm and a large age-range of children.

We believe that this paper represents critical steps towards an effective robot tutor for young children. Future work will incorporate a more interactive engagement between robot and child, such as incorporating reactions to the child's posture (Sanghvi et al. 2011), facial expressions (Baur et al. 2013) and speech (Bolaos et al. 2013; Cheng et al. 2014). A larger curriculum can also be employed that encompasses phonology, orthography, morphology, syntax and semantic meanings (Wolf et al. 2009), enabling a better personalization of the taught material. Developing the assessment Bayesian active-learning algorithm to these knowledge domains is another challenge we will address in the future.

To conclude, we have presented an integrated system that properly and efficiently assesses a child's word-reading skill during an engaging child-tablet-robot interaction, and uses the acquired knowledge to teach the child, in an empowering situation, new reading skills.

Acknowledgments

The authors acknowledge help and support of Jacqueline Kory in the development of the experimental setup. G.G. was supported by the Fulbright commission for Israel, the United States-Israel Educational Foundation. This research was supported by the National Science Foundation (NSF) under Grants CCF-1138986. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not represent the views of the NSF.

References

Astington, J. W., and Jenkins, J. M. 1999. A longitudinal study of the relation between language and theory-of-mind development. *Developmental psychology* 35(5):1311.

Baur, T.; Damian, I.; Lingenfelser, F.; Wagner, J.; and Andr, E. 2013. *NovA: Automated Analysis of Nonverbal Signals in Social Interactions*, volume 8212 of *Lecture Notes in Computer Science*. Springer International Publishing. book section 14, 160–171.

Bolaos, D.; Cole, R. A.; Ward, W. H.; Tindal, G. A.; Schwanenflugel, P. J.; and Kuhn, M. R. 2013. Automatic assessment of expressive oral reading. *Speech Communication* 55(2):221–236.

Breazeal, C., and Scassellati, B. 1999. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS '99. Proceedings. 1999 IEEE/RSJ International Conference on*, volume 2, 858–863 vol.2.

Brown, L.; Kerwin, R.; and Howard, A. M. 2013. Applying behavioral strategies for student engagement using a robotic educational agent. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 4360–4365.

Cakmak, M.; Chao, C.; and Thomaz, A. L. 2010. Designing interactions for robot active learners. *Autonomous Mental Development, IEEE Transactions on* 2(2):108–118.

Chambers, B.; Slavin, R. E.; Madden, N. A.; Abrami, P.; Logan, M. K.; and Gifford, R. 2011. Small-group, computer-assisted tutoring to improve reading outcomes for struggling first and second graders. *The Elementary School Journal* 111(4):625–640.

Cheng, J.; DAntilio, Y. Z.; Chen, X.; and Bernstein, J. 2014. Automatic assessment of the speech of young english learners. *ACL 2014* 12.

Fasola, J., and Mataric, M. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction* 2(2):3–32.

Fridin, M. 2014. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & Education* 70(0):53–64.

Jeong, G.-M.; Park, C.-W.; You, S.; and Ji, S.-H. 2014. A study on the education assistant system using smartphones and service robots for children. *Int J Adv Robot Syst* 11:71.

Kegel, C. A., and Bus, A. G. 2012. Online tutoring as a pivotal quality of web-based early literacy programs. *Journal of Educational Psychology* 104(1):182.

Kory, J. M.; Jeong, S.; and Breazeal, C. L. 2013. Robotic learning companions for early language development. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 71–72. ACM.

Leyzberg, D.; Spaulding, S.; and Scassellati, B. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 423–430.

MacKay, D. 1992. Information-based objective functions for active data selection. *Neural computation* 4(4):590–604.

Movellan, J.; Eckhardt, M.; Virnes, M.; and Rodriguez, A. 2009. Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 307–308.

Olney, A.; DMello, S.; Person, N.; Cade, W.; Hays, P.; Williams, C.; Lehman, B.; and Graesser, A. 2012. *Guru: A Computer Tutor That Models Expert Human Tutors*, volume 7315 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. book section 32, 256–261.

Polson, M. C., and Richardson, J. J. 2013. *Foundations of intelligent tutoring systems*. Psychology Press.

Saerbeck, M.; Schut, T.; Bartneck, C.; and Janse, M. D. 2010. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1613–1622. ACM.

Sanghvi, J.; Castellano, G.; Leite, I.; Pereira, A.; McOwan, P. W.; and Paiva, A. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, 305–311.

Setapen, A. 2012. *Creating Robotic Characters for Long-term Interaction.* Thesis.

Tanaka, F., and Matsuzoe, S. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1(1).

Tapus, A.; Maja, M.; and Scassellatti, B. 2007. The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine* 14(1).

VanLehn, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46(4):197–221.

Wise, B. W.; Ring, J.; and Olson, R. K. 2000. Individual differences in gains from computer-assisted remedial reading. *Journal of Experimental Child Psychology* 77(3):197–235.

Wolf, M.; Barzillai, M.; Gottwald, S.; Miller, L.; Spencer, K.; Norton, E.; Lovett, M.; and Morris, R. 2009. The raveo intervention: Connecting neuroscience to the classroom. *Mind, Brain, and Education* 3(2):84–93.