A Semantics-based Model for Predicting Children's Vocabulary

Ishaan Grover , **Hae Won Park** and **Cynthia Breazeal**
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, Massachusetts, 02139
{igrover, haewon, breazeal}@mit.edu

Abstract

Intelligent tutoring systems (ITS) provide educational benefits through one-on-one tutoring by assessing children's existing knowledge and providing tailored educational content. In the domain of language acquisition, several studies have shown that children often learn new words by forming semantic relationships with words they already know. In this paper, we present a model that uses word semantics (semantics-based model) to make inferences about a child's vocabulary from partial information about their existing vocabulary knowledge. We show that the proposed semantics-based model outperforms models that do not use word semantics (semantics-free models) on average. A subject-level analysis of results reveals that different models perform well for different children, thus motivating the need to combine predictions. To this end, we use two methods to combine predictions from semantics-based and semantics-free models and show that these methods yield better predictions of a child's vocabulary knowledge. Our results motivate the use of semantics-based models to assess children's vocabulary knowledge and build ITS that maximizes children's semantic understanding of words.

1 Introduction

Intelligent tutoring systems (ITS) provide educational benefits through one-on-one tutoring by assessing children's existing knowledge and providing tailored educational content [VanLehn, 2011]. In vocabulary learning tasks, most ITS come with a fixed curriculum (or a set of words) that the system attempts to teach without considering prior knowledge of the tutee. However, when human tutors interact one-to-one with a tutee, they first attempt to understand existing knowledge of the tutee (often from partial and noisy data).

Recent work in ITS for language learning has focused on modeling children's word-reading skills, pronunciation skills and affective states to provide personalized curricula [Park *et al.*, 2019; Spaulding *et al.*, 2018; Gordon *et al.*, 2016; Gordon and Breazeal, 2015]. However, much of language acquisition deals with understanding words in the context of other semantically related words. Studies have shown that children, as young as 3 years of age, often form categories among new objects using their shared semantic properties [Jones *et al.*, 1991]. During comprehension process, children retrieve previously learned vocabulary words in order to make new **conceptual associations** [Johnson *et al.*, 1982]. Furthermore, when children learn vocabulary along with semantics, they learn more words and their learning pace is faster [Wolf *et al.*, 2009; Dilek and Yürük, 2013]. It can be concluded from these studies that *(i)* children form a cognitive semantic representation for storage and retrieval from memory, and *(ii)* children learn better when taught words that are semantically related.

Even though learning through semantic associations forms the basis of many psycholinguistic theories of language acquisition, to the best of our knowledge, there hasn't been an attempt to create and experimentally validate a computational model that makes predictions about children's vocabulary using semantic associations between words. Since every child has a different vocabulary that can only be partially observed, one of the main challenges in building such a model is making inferences based on partial information while representing the assumptions of learning through semantics. To this end, we use Markov Random Field (MRF) models to probabilistically represent children's semantic knowledge and make inferences by capturing the assumptions of semantic learning. In the rest of this paper, we refer to this model as *semantics-based model* and models that don't take into account semantic relations between words as *semantics-free models*, such as frequency- and phonetics-based models.

The following example provides an intuition for the model presented in this paper. Consider a hypothetical case where an adult tutor tries to predict whether a child knows the target word "Jupiter". If it is identified that the child knows the words "Earth" and "Venus", the tutor's belief about knowledge of the target word becomes stronger. Now, if it is identified that the child does not know the word "Planet", this belief becomes weaker. Thus, when making inferences about a child's knowledge, the tutor first assumes that the words presented are semantically related to one another. This forms the tutor's *prior* knowledge. Further, the tutor assumes that the child stores words in memory using similar semantic relations. When the tutor obtains new information, s/he uses this information as *evidence* to update their *belief* about the

child's knowledge of the target word.

An ITS may use a similar reasoning to first observe the existing vocabulary of a child (partially) and then predict the probability of knowing other words that the child may or may not know. For example, a robotic educational companion participating in a story-telling task with children could observe their current vocabulary from the words they use in their stories [Park *et al.*, 2019]. Using models presented in this paper, the robot could then identify other words they may or may not know. Finally, using this knowledge, the robot could then formulate a strategy to enhance children's vocabulary. This task is further detailed in Future Work in section 9.

The main research questions addressed in this paper are presented as follows.

1.1 Research Questions

- **R1:** Given a partial observation of existing vocabulary of a child, can we build a model to predict whether s/he would know other semantically related words *(semantics-based model)*?
- **R2:** Can we use semantics-free models in conjunction with the semantics-based model to make better predictions about a child's existing vocabulary?

2 Related Work

In this work, we draw from and build upon two main research areas:

2.1 Cognitive Models for Vocabulary Acquisition

Most research in building cognitive models focuses on building models that learn word-to-concept mappings from data akin to how a child learns. For instance, Siskind proposed a model that used the principles of cross-situational learning to learn word-to-concept mappings [Siskind, 1996]. Yu et al. incorporated social cues like joint attention into a unified statistical learning framework for cross-situational observations [Yu and Ballard, 2007]. More recently, a model to build and incrementally grow a semantic network from utterance data was proposed to capture semantic relationships between words [Nematzadeh *et al.*, 2014].

However, when making predictions about a child's existing knowledge, often we do not have access to the data stream that the child was exposed to when they learned the words that exist in their current vocabulary. This motivates the development of models that make assumptions about children's existing vocabulary to make inferences about other words they might know.

2.2 Knowledge Prediction

Within ITS literature, researchers have proposed models for assessing children's reading skills and pronunciation. For instance, Gordon et al. [Gordon and Breazeal, 2015] presented a Bayesian active learning model to predict the probability of a child's ability to *read* a word. Spaulding et al. [Spaulding *et al.*, 2016] showed that incorporating affect information (smile and engagement) into a Bayesian knowledge tracing model outperforms traditional models for predicting a child's reading skill level. Researchers have also used Gaussian process regression to model children's *pronunciation skills* using a covariance function that is a weighted sum of semantic and phonemic similarity [Spaulding *et al.*, 2018].

However, there has been little work in predicting *word understanding* through semantics. Recently, researchers represented words using the Osgood semantic scale and word2vec embeddings and used mixed effect logistic regression to predict short- and long-term word acquisition in children [Nam *et al.*, 2017]. However, this model did not consider how semantic knowledge is represented and stored in children's memory.

3 Dataset and Terminology

We used the dataset from our work in [Park *et al.*, 2019]. We used data collected from the control (baseline) group, i.e., children who completed the *(i) Pre-test* and *(ii) Post-test* (3 months apart) without any technology intervention in between (such as in the experimental groups). The baseline group dataset included test results from 23 children between the ages of 5–6 from 12 kindergarten classrooms. Data from six children who completed the pre-test but not the post-test was excluded from our experiment, resulting a pair-wise dataset from 17 children (age $\mu = 5.39$, $\sigma = 0.48$; female 55.5%). The goal of the pre- and post-test was to sample words from a child's vocabulary. It is hard to measure a child's current vocabulary level since the set of words a child could know is very large. Thus, each child's linguistic level was assessed using a clinically evaluated vocabulary test called the Peabody Picture Vocabulary Test (PPVT) [Dunn and Dunn, 2007] which is one of the most widely used tests to measure a child's vocabulary level. The format of the test involves asking a child to select one of four possible pictures that are related to the target word being assessed. The words presented in the test are such that they align with the vocabulary level of the child. The test ends when the words presented are above the vocabulary level of the child, and the child is consistently unable to answer the questions. After repeating this process for multiple words in the test, we are able to obtain a dataset containing a sample of words that a child knows and doesn't know within the child's linguistic abilities. After three months, a post-test PPVT was performed to again sample words. A period of three months minimized the effect of any short-term storage of word associations encountered during pre-test. The set of target words assessed during the post-test (PPVT-4, Form B) were completely different from those used in the pre-test (PPVT-4, Form A). One common word was excluded from the post-test and included in the pre-test only.

The data collected from the experiment is summarized as follows:

- Pre-test: Set of words W_{pre} and information about whether or not they know those words.
- Post-test: Set of words W_{post} and information about whether or not they know those words. Moreover, $W_{post} \cap W_{pre} = \emptyset$.

4 Preliminaries

A Markov Random Field (MRF) is an undirected graphical model defined by graph $G_{mrf} = (V_{mrf}, E_{mrf})$ and a set of random variables X where vertices $V_{mrf} = \{v_1, v_2, v_3...v_n\}$ correspond to random variables $X = \{X_1, X_2, X_3...X_n\}$. An edge e_{ij} between nodes X_i and X_j captures the notion of dependence or interactions between nodes. This dependence is numerically represented by a potential function $\phi(\mathbf{x})$ which may be defined for a pair of nodes or a clique (c) in the graph. When it is defined for pairs of nodes, the MRF is called a pairwise MRF. The potential functions are often represented as energy functions and then transformed into probabilities by adopting Gibbs distribution. The lower the energy of a clique, the higher is its potential and thus higher the probability. Thus, we have:

$$P(X_1, X_2...X_n) = \frac{1}{Z} \prod_{C} \phi_c(\mathbf{x}_c) \tag{1}$$

$$Z = \sum_{\mathbf{x}} \prod_{C} \phi_c(\mathbf{x}_c) \tag{2}$$

$$\phi_c(\mathbf{x}_c) = e^{-E(\mathbf{x}_c)} \tag{3}$$

where,

- C is the set of all maximal cliques in the graph.
- $\phi_c(\mathbf{x}_c)$ is the potential function associated with clique c.
- $\phi_c(\mathbf{x}_c) \geq 0$.
- $E(\mathbf{x}_c)$ is the energy function associated with the clique c.
- Z is the normalizing constant or the partition function.

A common algorithm to perform inference on MRF models is Belief Propagation (BP). In this paper, we use the sum-product variant of BP to compute the marginal probability distribution of nodes.

5 Models for Knowledge Prediction

The main contribution of this paper is a model that makes inferences assuming the semantic model of vocabulary acquisition for the learner. The fundamental assumption made by this model in order to make inferences is the following:

A1. Children learn words by forming semantic relations with existing words that they know. Thus, if it is observed that a child knows a word, it is likely that the child knows words that are semantically related to the given word. On the other hand, if it is observed that the child does not know a given word, it is likely that the child does not know words that are semantically related to the given word.

Given the experimental design, we now restate the first research question more specifically in terms of the data collected from the experiment:

R1. Given a child's knowledge about words assessed in the pre-test, can we build a model to predict whether s/he would know other semantically related words as assessed in the post-test (*semantics-based model*)?

5.1 Semantics-based Model - MRF

R1 as an Inference Problem

In order to answer R1, we formulate it as an inference problem on undirected graphical models (UGMs). Broadly, we first build a semantic network where the nodes represent words and edges represent semantic relationship between the words. Then, every node in the semantic network is mapped to a random variable of an MRF representing the probability of knowing or not knowing the word. Finally, we perform inference to find the marginal probabilities of nodes (or word) after using words in W_{pre} as evidence.

Measure of Semantic Similarity

A common measure of semantic distance between words is to take the cosine distance between the word embeddings of two words. Here, we define semantic distance between two words as the cosine distance between their pre-trained common crawl GloVe word vectors (300 dimensional) [Pennington *et al.*, 2014]. Two words with vector representations v_1 and v_2 are said to be semantically similar if *cos*(v_1, v_2) $\geq \epsilon$ (after manually testing different values of semantic similarity on different words, we set $\epsilon = 0.6$ for this study).

Building Semantic Network

When estimating a child's vocabulary from partial data, a human tutor often has some prior knowledge about a rough estimate of the kinds of words a child might know and how they are semantically related to each other. Hence, to represent this belief about a child's knowledge computationally, we build a semantic network using the first thousand words W_{list} [Fry, 1980] that should be taught to children and are now increasingly adopted by many schools along with words in W_{pre} and W_{post}. Let this graph be called $G_{semantics}$ = (V, E) such that V = $W_{pre} \cup W_{post} \cup W_{list}$. Set of edges E of the graph are computed using pairwise comparisons between nodes to check for semantic similarity of words, i.e., nodes v_1 and v_2 are connected by edge e_{12} if and only if v_1 and v_2 are semantically similar. Building this graph runs in $O(|V|^2)$ which is computationally acceptable in our case ($|V| \approx 1,400$). We further define two words w_1 and w_2 to be *semantically related* if there exists a path between them in $G_{semantics}$. Thus, a node is semantically related to another node through a path of semantically similar nodes. We use GloVe word vectors [Pennington *et al.*, 2014] to build the graph instead of other publicly available semantic graphs due to the flexibility of computing semantic similarity between any given pair of words and the ability to use specific words in the semantic graph that are expected to be within the linguistic abilities of children.

Semantic Network to MRF

The graph structure of a semantic network is similar to that of an MRF. We create a graph $G_{mrf} = (V_{mrf}, E_{mrf})$ from semantic network $G_{semantics} = (V, E)$ such that $|V_{mrf}| = |V|$ and $|E_{mrf}| = |E|$. Every node V_i in $G_{semantics}$ is mapped to a Bernoulli random variable $X_{i,mrf}$ in G_{mrf} which represents the distribution of whether or not a child knows the corresponding word in $G_{semantics}$. Two nodes $V_{i,mrf}$ and $V_{j,mrf}$ are connected in G_{mrf} if and only if their corresponding nodes are connected in $G_{semantics}$.

Potential Functions

Since we only have a measure of semantic similarity between pairs of words, we use pairwise potential functions to capture the notion of how two words are semantically similar to each other. In order to represent the previously stated assumption $A1$, potential functions must further be associative in nature. That is, they should favor neighboring nodes to have the same label (knowing the word or not knowing the word) giving rise to a model that is both pairwise and associative in nature. We define the energy function as follows: if X_i corresponds to the word w_i and X_j corresponds to the word w_j in graph $G_{semantics}$, and $s(w_i, w_j)$ gives the semantic similarity[1] between words w_i and w_j, we define the energy of neighboring nodes having the same label as:

$$E(X_i, X_j) = 1 - s(w_j, w_j) \tag{4}$$

and energy of neighboring nodes having a different label as:

$$E(X_i, X_j) = s(w_j, w_j) \tag{5}$$

Thus, the higher the semantic similarity between two nodes, the lower will be the energy associated with the pair of words. This results in the final pairwise potential function:

$$\phi(X_i, X_j) = \begin{bmatrix} e^{-(1-s(w_i,w_j))} & e^{-s(w_i,w_j)} \\ e^{-s(w_i,w_j)} & e^{-(1-s(w_i,w_j))} \end{bmatrix} \tag{6}$$

Inference

Given the structure and potential functions of the MRF, we now perform BP using words in W_{pre} as evidence to find the marginal probabilities of words in W_{post}.

5.2 Semantics-based Baseline Models

The proposed graphical model (MRF) has two key components: *(i)* A graph representing semantic relationships and *(ii)* the ability to perform inference. We now discuss models that use semantics but lack some of the features of the model presented. We use these models or algorithms as baselines to later compare the performance of the graphical model.

GloVe Nearest Neighbor

To predict whether a child knows a given word or not, we find the word in W_{pre} that has the highest measure of semantic similarity with the given word and assign the label corresponding to the word. Hence, this model neither has a semantic graph representation nor does it have the ability to perform inference.

Semantic Network Nearest Neighbor

Another strategy to make predictions is to use the graph structure of the semantic network, but not cast it as an MRF or perform inference. Instead, for a given target word for which a prediction is to be made, we perform Breadth First Search (BFS) using the word as the source, and find the label associated with the word in W_{pre} that has the shortest path distance from the target word.

[1]When computing semantic similarity, we use the lemmatized form of words. This is done so that different forms of words don't affect their semantic similarity. For example, semantic distance between "cats" and "dogs" should be the same as that between "cat" and "dog". We use the lemmatized representation for all models presented in the paper except when specified.

5.3 Semantics-free Models

In this section, we describe models that do not consider word semantics but may be used to make predictions about a child's vocabulary knowledge using some prior ground truth about what the child knows and doesn't know. Namely, we discuss frequency-based and phonetics-based models.

Frequency-based Model

The theory of *incidental learning* posits that language learners often "pick up" new words through reading, listening and conversational activities [Hulstijn and others, 2003]. Incidental vocabulary acquisition has a direct link with frequency of exposure - the higher the frequency of exposure to a word, the higher will be the probability of a child committing it to memory [Teng, 2016].

We use the SUBTLEXus database (74,286 word forms) as a source of word frequency counts of different words used in spoken English language [Brysbaert and New, 2009]. We use the zipf scale measure of word frequency counts instead of raw frequency counts. The zipf scale converts word frequencies (per billion words) into a log-based scale with values between 1-7 and is independent of the size of the corpus used [Van Heuven *et al.*, 2014]. Then for each child, we train a personalized logistic regression model using zipf score of words in W_{pre} as training data and whether or not a child got the words correct as training labels (binary classification). Since the dataset used to train the model had a class imbalance (i.e, there were more instances of words known than words unknown), we weighted the classes inversely proportional to their frequency in the training set.

The weight w_i for class i is given by:

$$w_i = \frac{n}{kn_i} \tag{7}$$

where n is the total number of data points, k is the number of classes (here, $k = 2$) and n_i is the number of instances that have label i. In this way, the model is able to learn the kinds of words that a child knows based on frequency of word occurrence in everyday language use.

Phonetics-based Model

Apart from the frequency of word exposure, it is also studied that children use phonological information of words they already know to learn new words [Edwards *et al.*, 2004]. Thus, for a given target word, the nearest phonologically similar word in the child's vocabulary would give information about whether the child had the phonological knowledge to have learned the given target word. A common measure of phonological distance between words is Levenshtein distance which measures the edit-distance of insertion, deletion and replacement of characters between two words [Sanders and Chin, 2009].

Since we can only observe a child's vocabulary partially, we use words in W_{pre} as a measure of the child's phonological knowledge. Thus, to make a prediction for a given word, we find the nearest word in W_{pre} (using the Levenshtein distance metric) and assign the label of the nearest word (we use a fixed cost of 1 for each operation). For example, if a child knows the words "cat" and "rat", it would mean that the child

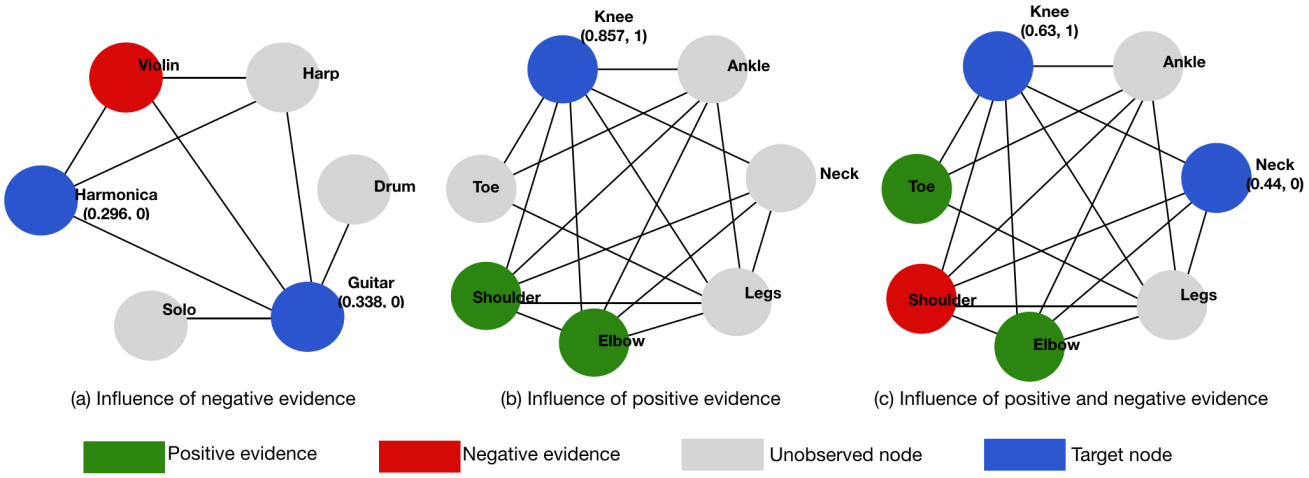

Figure 1: Predictions made by MRF for different subjects. The three graphs in (a), (b) and (c) are subgraphs of graph G_{mrf}. The positive and negative evidence are labels corresponding to words in W_{pre}. Predictions and true labels for target nodes are shown as (prediction, true label). (a) shows the influence of negative evidence to make predictions. (b) shows the influence of positive evidence to make predictions (c) shows how different observations (negative and positive) impact the final predictions for target nodes.

has the phonological knowledge to also have learned the word "bat". This strategy allows the algorithm to capture the notion of retention of words based on their phonological knowledge using the Levenshtein distance as a proxy measure of phonological word similarity [2].

6 Ensemble Methods

Different cognitive theories posit different ways of how children acquire new words and no one theory alone explains how children learn words. Each model discussed in the previous section is based on a different cognitive theory of learning. In this section, we propose two strategies of combining predictions from different models to evaluate if we can make better predictions by combining the predictive capabilities of each model (**R2**).

6.1 Conditional Independence

We wish to estimate the probability of a child knowing a given word w, and have two separate models m_1 and m_2 that estimate the probability of the child knowing the word. Thus, we are given $p(w|m_1)$ and $p(w|m_2)$ and want to find $p(w|m_1, m_2)$. If we assume the two models to be conditionally independent sources, then the probability $p(w|m_1, m_2)$ using the Bayes optimal method to combine distributions is given by [Bailer-Jones and Smith, 2011]:

$$p(w|m_1, m_2) \propto p(w|m_1)p(w|m_2) \tag{8}$$

6.2 Mixture of Distributions

Another common method of combining probabilities is to create a new distribution that is a mixture of two distributions (weighted sum). Since there is no prior informing which of the two distributions should be given a higher weight, we assume an equal weight (0.5) for each of the distributions and compute the posterior by taking the weighted sum of distributions from the two models.

[2]When computing levenshtein distance, we use the given word forms instead of their lemmatized forms because lemmatization can sometimes change the phonological structure of a word.

7 Evaluation

To answer R1 and R2, we report *(i)* area under the precision-recall curve for *(ii)* words in W_{post} that are semantically related to words in W_{pre}. The two decisions are justified as follows:

Selection of words. For any given child, not all words in W_{post} are semantically related to words in W_{pre}. This occurs in the case where $G_{semantics}$ is a disjoint graph and there exists a word $w_{i,post}$ in an independent subgraph in which there exist no words from W_{pre}. For example, if a human tutor only knows about a child's knowledge of words related to the solar system (earth, mars, moon, etc), they cannot make predictions about words that are related to computers (laptop, keyboard, screen, etc) using only word semantics. Semantics-free models however have the ability to make predictions about any given word irrespective of their semantic relations. Thus, we compare the models only on words $W_{post,selected}$ in W_{post} that have some path to words in W_{pre}.

Area under the precision-recall curve. A common method of evaluating probabilistic binary classifiers is area under the curve (AUC) of the receiver operating curve (ROC) as it tries to balance true positive and false positive rates by considering a number of different thresholds. The given dataset had a class imbalance (72%). In such cases, area under ROC can provide a skewed picture of the model's performance. Hence, we compute area under the precision-recall curve which is a more informative metric for datasets with class imbalances [Davis and Goadrich, 2006].

8 Results and Analysis

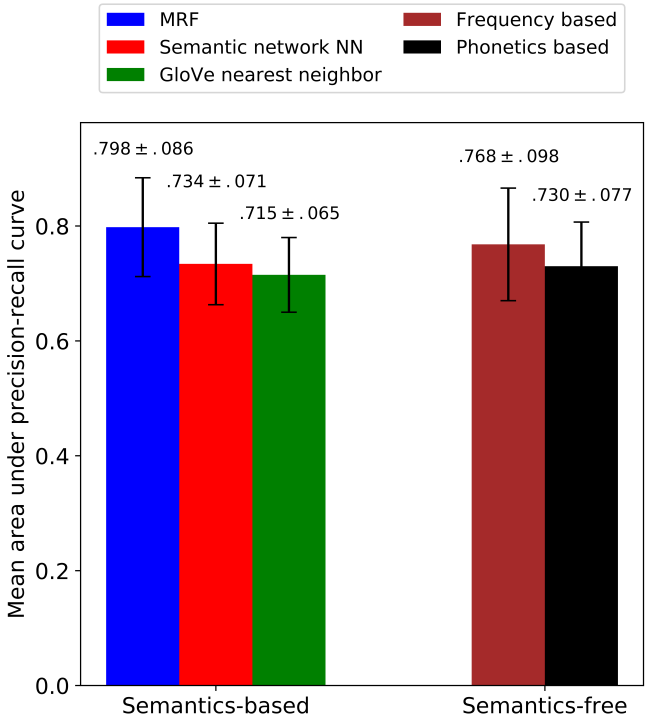

Figure 2: Mean area under precision-recall curve for all subjects.

Figure 1 shows subgraphs of G_{mrf} and predictions made by the MRF model. Figure 2 shows the mean area under the precision-recall curve for each of the models. We find that among semantics-based models, MRF ($AUC \approx 0.80$) has the highest mean AUC. Further, we find that the semantic network representation ($AUC \approx 0.73$) has a higher predictive power when compared to a model that finds the nearest semantically similar neighbor ($AUC \approx 0.72$). More interestingly, the results show that the real advantage of an MRF comes from observing nodes and performing inference. The significant increase in AUC from ≈ 0.73 (nearest neighbor in semantic network) to ≈ 0.80 (MRF) shows that the posterior probability of knowing a word is determined by observations of all nodes in W_{pre} in the graph and not just the nearest neighbor. Between the semantics-free models, the frequency-based model performed significantly better ($AUC \approx .77$) than the phonetics-based model ($AUC \approx 0.73$). When comparing the frequency-based model with MRF, we find that MRF has a better performance in predicting words in $W_{post,selected}$. It is interesting to note that MRF is able to perform well using inferences based on word similarities alone without any information about how often a child might have been exposed to a given target word.

8.1 Subject-level Analysis

Between MRF and the frequency-based model, a subject-level analysis allows us to find subjects for which either of the models is better in predicting knowledge of words in $W_{post,selected}$. Table 1 shows that out of 17 subjects, MRF performed better than the frequency-based model on 10 subjects. Moreover, we observe that the mean AUC scores on subjects where MRF performs better is higher than the frequency-based model. More concretely, on the set of subjects where MRF dominates, $AUC_{mrf} \approx 0.82$ and $AUC_{freq} \approx 0.736$, while on the set of subjects where the frequency-based model performs better, $AUC_{freq} \approx 0.81$ and $AUC_{mrf} \approx 0.77$. Figure 3 shows the difference $AUC_{mrf} - AUC_{freq}$ per subject. Here, we further see that there is a distinct set of subjects where each of the models performs significantly better than the other.

Thus, in reference to the first research question R1, the aforementioned results show that MRF is effective in making predictions about semantically related words in post-test using words in the pre-test as observations.

Condition	# subjects	AUC_{mrf}	AUC_{freq}
$MRF > Freq$	**10**	$\mathbf{.82 \pm .096}$	$.736 \pm .108$
$MRF < Freq$	7	$.765 \pm .053$	$\mathbf{.814 \pm .055}$

Table 1: Subject-level analysis of mean area under precision-recall curve.

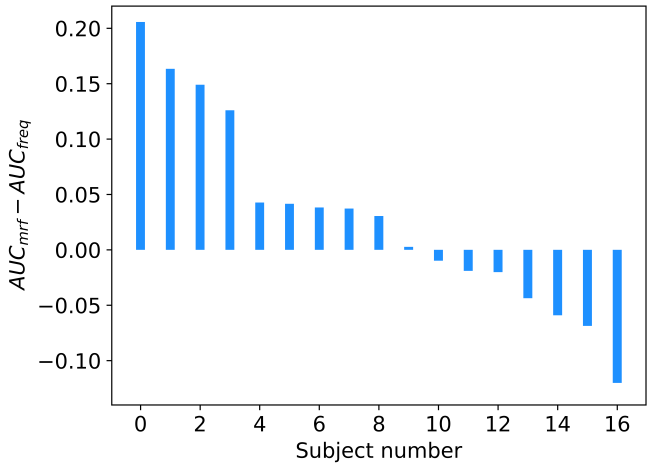

Figure 3: Subject-level differences between AUC_{mrf} and AUC_{freq}.

Model	Mean AUC	Std
MRF	0.798	0.086
frequency-based	0.768	0.098
Combined (conditional independence)	0.802	0.088
Combined (mixture of distributions)	**0.803**	**0.088**
Best model (per child)	0.818	0.082

Table 2: Area under precision-recall curve after combining predictions.

8.2 Combining Predictions

Since each model makes assumptions according to different psycholinguistic theories of learning and the fact that both MRF and the frequency-based model perform well of different sets of subjects, we evaluate if combining predictions increases the performance of either of the models. Table 2 shows that there is a slight improvement in prediction performance for MRF ($\approx 0.5\%$) and a greater improvement in performance for the frequency-based model ($\approx 3.5\%$). Both strategies of combining predictions give similar gains in improvement. This improvement is seen even though the set of children where the two models perform well are mutually exclusive and the difference in performance is significant, as seen in Figure 3. These results show that combining two models that are based on different theories of learning can help improve prediction performance.

Limitations

It is important to note that the predictions were combined using uniform priors for both methods. While the strategies to combine predictions work, the improvement in not significant (in the case of MRF) because of the absence of informative priors. Table 2 shows the results when the best model (MRF or frequency-based) is used for making predictions for each child – i.e, for each child, we assign a weight of 1.0 to the model that performs better and a weight of 0.0 to the other. This method shows that if we could accurately pre-determine which model would perform best for each child, the best achievable AUC ≈ 0.82. Therefore, given that each model performs well on different subjects and the fact that predictions can be combined using the aforementioned strategies, methods (or heuristics) to pre-determine the weights (*priors*) for each model per subject would further help make better predictions about a child's knowledge.

Thus, in reference to the research question R2, results show that it is possible to make better predictions by combining the predictions of individual models (MRF and frequency-based) with a caveat that more informative priors would significantly help in combining predictions to achieve better performance.

9 Future Work

The models presented in this paper allow us to make predictions about a child's vocabulary. Future work will investigate how these models can be implemented on an educational robot companion to enhance children's vocabulary to extend our previous work [Park *et al.*, 2019]. For example, in a setting where a child and a robot take turns telling each other stories, the robot's goal would be to enhance a child's vocabulary by asking questions about specific semantically related words as the child narrates their story. Using MRF, the robot would first determine the probability of knowing different words (from knowledge about words they use in their stories). The robot would then choose to ask questions about words it is certain that the child knows. Our hypothesis is that this strategy would provide encouragement and positive engagement to the child. On the other hand, the robot could also ask questions that it is certain the child doesn't know and then attempt to teach the meaning of those words. This strategy would directly enhance children's knowledge. For either strategies, the models presented in this paper are a prerequisite. Thus, in future work, the robot would learn to adapt it's strategy based on children's engagement and existing knowledge to maximize their learning.

10 Conclusion

Different psycholinguistic theories posit different ways in which children learn new words. In this paper, we present models for predicting children's vocabulary from partial knowledge of their existing vocabularies and ground assumptions made by each model in different theories. More concretely, we present MRF (based on semantic association between words), a frequency-based model (based on theory of incidental learning) and a phonetics-based model. We further show how predictions from different models such as frequency-based model and MRF can be combined. Using data from 17 subjects, we experimentally show that MRF can effectively predict children's vocabulary. We further show that the ensemble of MRF and frequency-based model can further improve prediction performance of individual models. These results motivate the use of semantics-based models in ITS to assess children's knowledge.

Acknowledgements

This work was supported by a National Science Foundation grant, IIS-1734443. The authors would like to sincerely thank Dr. Goren Gordon, Samuel Spaulding, Abhimanyu Dubey, Spandan Madan, Adam Haar Horowitz and reviewers for their valuable comments and feedback on this work.

References

[Bailer-Jones and Smith, 2011] C Bailer-Jones and K Smith. Combining probabilities. *Data Processing and Analysis Consortium (DPAS)*, 2011.

[Brysbaert and New, 2009] Marc Brysbaert and Boris New. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–90, 11 2009.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[Dilek and Yürük, 2013] Yeşim Dilek and Nurcihan Yürük. Using semantic mapping technique in vocabulary teaching at pre-intermediate level. *Procedia-Social and Behavioral Sciences*, 70:1531–1544, 2013.

[Dunn and Dunn, 2007] Lloyd M Dunn and Douglas M Dunn. *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments, 2007.

[Edwards *et al.*, 2004] Jan Edwards, Mary E Beckman, and Benjamin Munson. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of speech, language, and hearing research*, 2004.

[Fry, 1980] Edward Fry. The new instant word list. *The Reading Teacher*, 34(3):284–289, 1980.

[Gordon and Breazeal, 2015] Goren Gordon and Cynthia Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *AAAI*, pages 1343–1349, 2015.

[Gordon *et al.*, 2016] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *AAAI*, pages 3951–3957, 2016.

[Hulstijn and others, 2003] Jan H Hulstijn et al. Incidental and intentional learning. *The handbook of second language acquisition*, pages 349–381, 2003.

[Johnson *et al.*, 1982] Dale D Johnson, Susan Toms-Bronowski, and Susan D Pittelman. An investigation of the effectiveness of semantic mapping and semantic feature analysis with intermediate grade level children (program report 83-3). madison: University of wisconsin. *Wisconsin Center for Educational Research*, 1982.

[Jones *et al.*, 1991] Susan S Jones, Linda B Smith, and Barbara Landau. Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516, 1991.

[Nam *et al.*, 2017] SungJin Nam, Gwen Frishkoff, and Kevyn Collins-Thompson. Predicting short-and long-term vocabulary learning via semantic features of partial word knowledge. *Ann Arbor*, 1001:48109, 2017.

[Nematzadeh *et al.*, 2014] Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. A cognitive model of semantic network learning. In *EMNLP*, pages 244–254, 2014.

[Park *et al.*, 2019] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *AAAI*, 2019.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Sanders and Chin, 2009] Nathan C Sanders and Steven B Chin. Phonological distance measures. *Journal of Quantitative Linguistics*, 16(1):96–114, 2009.

[Siskind, 1996] Jeffrey Mark Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, 1996.

[Spaulding *et al.*, 2016] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[Spaulding *et al.*, 2018] Samuel Spaulding, Huili Chen, Safinah Ali, Michael Kulinski, and Cynthia Breazeal. A social robot system for modeling children's word pronunciation: Socially interactive agents track. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1658–1666, 2018.

[Teng, 2016] Feng Teng. The effects of word exposure frequency on incidental learning of the depth of vocabulary knowledge. *GEMA Online® Journal of Language Studies*, 16(3), 2016.

[Van Heuven *et al.*, 2014] Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190, 2014.

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

[Wolf *et al.*, 2009] Maryanne Wolf, Mirit Barzillai, Stephanie Gottwald, Lynne Miller, Kathleen Spencer, Elizabeth Norton, Maureen Lovett, and Robin Morris. The rave-o intervention: Connecting neuroscience to the classroom. *Mind, Brain, and Education*, 3(2):84–93, 2009.

[Yu and Ballard, 2007] Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.