Examining Early Language Acquisition in LSTMs and Children using Semantic Networks

Jonathan Lu (jxl219@nyu.edu)
Center for Data Science, New York University, 60 5th Ave, New York, NY 10011

Isabel Kerber (ick225@nyu.edu)
Department of Psychology, New York University, 6 Washington Place, New York, NY 10003

Peiling Jiang (pj787@nyu.edu)
Department of Psychology, New York University, 6 Washington Place, New York, NY 10003

Abstract

Children's language learning progresses impressively rapidly and efficiently. In the area of neural networks, Long short-term memory (LSTM) networks have had large successes on a variety of problems like speech recognition, language modeling, or translation. We compared language acquisition in an LSTM to that of children. Using semantic networks, we compared the vocabulary development of an LSTM trained for different amounts of time to that of children between the ages of 16 to 36 months. Further, we tested whether vocabulary growth follows similar principles as in children, specifically the principle of preferential attachment (Hills, Maouene, Maouene, Sheya, & Smith, 2009). Our results show that the vocabulary of the LSTM grows rapidly. While there are similar structures in the semantic networks between the LSTM and children early on, they differ over the course of development.

Keywords: Long short-term memory; word embedding; language acquisition

Introduction

Learning language is one of the most impressive human abilities and understanding vocabulary development can support our understanding of cognitive and linguistic development more generally. Children's acquisition of language starts very early and progresses rapidly. Given an average vocabulary of more than 60,000 words in 17-year-olds, children should learn about 10 new words per day from their first birthday on (Bloom, 2001; Miller & Gildea, 1987). Importantly, children can grasp a word's meaning without any training or feedback and their language-learning capacities improve over the course of their development (Taatgen & Anderson, 2002; Marcus, 1993).

Language allows humans to reason about a wide range of physical and psychological situations (Carey, 2009) and learn more quickly by relating new concepts to existing ones (Lupyan & Bergen, 2016). Understanding language acquisition is an important building block for building machines that learn and think like humans.

Long short-term memory (LSTM) networks, a special kind of recurrent neural network, have had large successes on a variety of problems including speech recognition, sequence learning, language modeling, and translation. What makes them special is their ability to learn long-term dependencies. Introduced by Hochreiter and Schmidhuber (1997), they were refined and popularized in following works. In this project, we want to compare children's language acquisition to that of an LSTM model. For data on children's language development, we turned to the MacArthur-Bates Communicative Development Inventories (CDI) (Fenson et al., 1993, 2007), a widely used family of parent-report instruments for data-gathering about early language acquisition. Based on CDI, Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017) introduced a structured database of children's vocabulary knowledge.

We focused on children's vocabulary growth and studied language acquisition measured as word production according to parental reports. To mimic children's language acquisition environment in an LSTM, we trained the network on various sources of famous children's books, films, and television shows. Further, we varied the amount of training for the network, which also influences the amount of exposure to the input data that the network can attain (see section LSTM Modeling). We then extracted the vocabulary size of each model by assessing their overall word production, in analogy to the production index in the CDI. To model the language semantics, we turned to networks, as they provide intuitive and useful representations of semantic knowledge.

Further, we wanted to test whether the networks grow according to the principle of preferential attachment (Steyvers & Tenenbaum, 2005). This form of growth can be observed in networks that adhere to a power-law distribution (or small-world structure), where a small but significant number of nodes are connected to a very large number of other nodes. More specifically, the probability that a node in the network connects with k other nodes is proportional to k^{γ}, where γ is a constant, called scaling parameter. Such scale-free structures can be observed in the World Wide Web or in the neural network of the worm Caenorhabditis elegans. This process generates scale-free (or power-law) degree distributions, (Pastor-Satorras, Smith, & Solé, 2003), and has been suggested as a growth process for semantic networks (Steyvers & Tenenbaum, 2005). Steyvers and Tenenbaum (2005) tested the hypothesis that the psychological process that underlies early noun learning follows the principle of preferential attachment. According to this principle, nouns that are learned ear-

liest in the network should show proportionally more connections at later stages of development. Using several measures of age of acquisition, their data showed a pattern of higher semantic connectivity in an associative network for early nouns than for later nouns. An analysis by (Hills et al., 2009) supported this pattern in children aging 16 to 30 months. This result could have profound implications for our understanding of word learning, as it suggests that growing expertise in one domain selects the entry of new information. How does it compare to word learning in an LSTM?

In the following, we describe how our LSTM was trained and data for analysis was obtained.

Data Understanding

We first collected and preprocessed the data for the model training. Unlike most training of LSTM models using books or articles, we collected several scripts for famous children's films and television shows, e.g., the conversational scripts of Mulan, Peter Pan, and Finding Nemo, as our text data. We used scripts of children's shows to best replicate the true language acquisition environment in their early ages as young children learn from what is spoken around them (Mehler et al., 2000) while written books or articles may otherwise better represent the expression habits of the author. Also, rich media, i.e. TV, film, and radio, themselves play an important role in the early stages of language development (Rice, 1983). All the data used for this paper can be found at https://github.com/jonathanlu02/lexical-classes-lstm/tree/main/data.

We then cleaned, preprocessed, and organized data into sequences of tokens for the model training. Some of our preprocessing includes:

1. **Expand contractions** - Contractions are shortened words or phrases in either oral or written language. Some letters and sounds are removed to create these contractions, e.g., in English, one of the vowels is usually removed, like *don't* contracted from *do not*. To help our model leverage standardized text, we first converted all contractions to their expanded version, e.g., converting *I'm* to *I am*.

2. **Remove non-alphabetical characters** - Here we intended to mimic the conversational language environment where children learn by listening and speaking. Thus, only words consisting of alphabets are meaningful for the modeling. We removed all non-alphabetical characters, including special characters like periods and commas, as well as digits.

3. **Lemmatization** - We want to always use the base form of the word to perform the model training as well as prediction. Thus, we remove the affixes of the words to get the root words, e.g. converting *making* to *make*, which is always lexicographically correct compared to the root stem obtained from stemming.

After cleaning and preprocessing the text, we got a long list of tokens of 243442 words with 11892 unique words. We then organized these tokens into sequences. We divided the sequences into 50 input words and 1 output word, and iterated through the whole preprocessed data resulting in 243391 sequences. The sequences data were stored as a text file with each line being one sequence, and for reference and can be accessed from the same place of the code. (See code availability section.)

LSTM Modeling

Initially, we proposed to use a bag of words to represent every word and then use cosine similarity to match every word between documents. That is, given a corpus of documents, we assign a frequency to how often every word appears. The problem we encountered was sparsity, where there were many uncommon words that didn't appear in most documents. This leads to performance issues when training the model. We also considered one-hot encoding each word. But similarly to bag of words, this technique created sparsity issues. Hence, we decided to use word embeddings, which assigns every word a vector format while also giving the word a sense of direction. One drawback to word embeddings is its inability to handle unknown or out-of-vocabulary words (Aylien News API, 2020). This tends to be a problem because you are forced to use a random vector, which is not ideal. In human cognition, if a child encounters a word they never seen before, their first thought is not to assign the word a random meaning. When a new word is encountered, they may be able to decipher the general meaning of this word through its context. Therefore word embeddings may struggle in the field of human cognition. But its ability to assign a direction to a word outweighs the drawbacks, and it is still the one of the best ways to represent words using current technology. This was implemented in the first layer (Embedding Layer) of the model. There are pre-trained word embedding models one can find for public use such as GloVe or Word2Vec (Mujtaba, 2020). These pre-trained models are trained on much larger databases and as a result pick up more semantic meaning. Hence, they typically lead to lower training errors and faster training time. But since we are modelling children's word development from scratch, using a pre-trained word embedding would be counter-intuitive. A typical word vector embedding is seen in Figure 1.

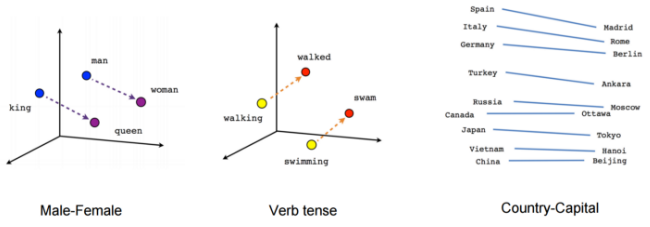

Figure 1: Word embeddings from (Soni, 2020)

After the Embedding layer, the data gets transferred to the LSTM layer. In order to best define our problem, we

need to use a model that is able to not only generate words that children mimic, but also provide context behind those words. Thus we considered using RNNs in our baseline model. Specifically a long short-term memory network, a type of RNN, is best suited for this task. It differs from the traditional RNN in the sense that it deals with the vanishing/exploding gradient problem where model weights diverge and become unstable during training, making it difficult to learn sequences of data.

Using an LSTM for this problem does come with some notable drawbacks, however:

1. **Memory Bandwidth Bounded due to Lack of Parallelization** - Unlike some other neural networks, LSTMs cannot be run in parallel, as each hidden state and cell state has to be computed before the next hidden state and cell state can be computed. LSTMs require four linear layers per cell to run for each sequence time-step. Linear layers require a large amount of memory bandwidth to be computed (Culurciello, 2018). During training, our memory usage was bottlenecking and being used at its maximum, not allowing GPU/CPU usage to catch up. Due to hardware limitations, we did not have as many layers in the model as we hoped.

2. **Overfitting** - The units in a network depend on the other units in the network. They try to minimize the loss function in the most efficient way possible, even if it means fitting the data perfectly. As a result, they may start to learn the training data and cause overfitting. Common methods to prevent overfitting include early stop, parameter regularization, and dropout. We will employ dropout layers in our model to prevent overfitting.

3. **Perfect memory retention** - Unlike LSTMs, humans don't memorize things perfectly. Rather, our memories are deeply embedded in context. According to temporal context models of memory (Howard & Kahana, 2002), learning consists of binding items to a context representation, which consists of both external elements like the place and setting, but also internal elements like the cognitive and emotional state of the individual. Based on the importance of experiences, like their association with punishment or reward, certain experiences attract attention and are processed preferentially (Talmi, Lohnas, & Daw, 2019).In contrast, the only context that LSTMs receive is the words in the data. Therefore, we used a bidirectional LSTM (BiLSTM) in our final model to give us slightly more context than a regular, one-directional (left to right) LSTM.

LSTMs are notoriously easy to overfit. In between each LSTM/Dense Layer, we incorporated a layer for dropout. Dropout mean that a unit from the neural network is temporarily removed from the network. The idea is that each hidden unit in a network must work with a randomly chosen sample of other units. Thereby, the hidden units are more robust without relying on other hidden units to correct its mistakes, and thus reduces overfitting. The paper by (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) explains this well: "In a standard neural network, the derivative received by each parameter tells it how it should change so the final loss function is reduced, given what all other units are doing." Therefore, units may change in a way that they fix up the mistakes of the other units. This may lead to complex co-adaptations. This in turn leads to overfitting because these co-adaptations do not generalize to unseen data Srivastava et al. (2014). The data we used only consisted of certain popular children's books and scripts. It would not make sense for a model to generate only words or sequences from these select media. In human cognition, children will read more books over time, including different books than the ones we selected. Children also develop semantic cognition through other sources, like conversations between parents. Conversations between characters in books and movies can be very different from actual conversations between people. Unfortunately, modeling such input was not possible. Conversely, too much dropout will cause underfitting, as the model will unable to learn effectively. Hence the importance of creating a model that generalizes well to unseen data, which can be achieved through drop-out. In the next section, we will analyze the results of our model.

Our final proposed model is shown in Figure 2.

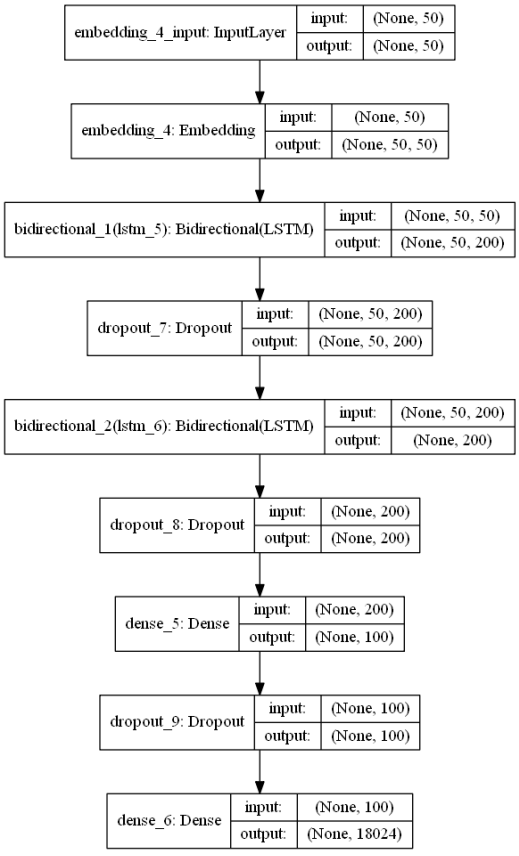

Figure 2: Final proposed model.

Model Results

In our baseline model (no dropout, regular LSTM), we experienced severe overfitting. After running the data across the model for 100 epochs, the training accuracy increased but the validation accuracy **decreased** as shown in Figure 3. Similarly, the training loss decreased while the validation loss **increased**.

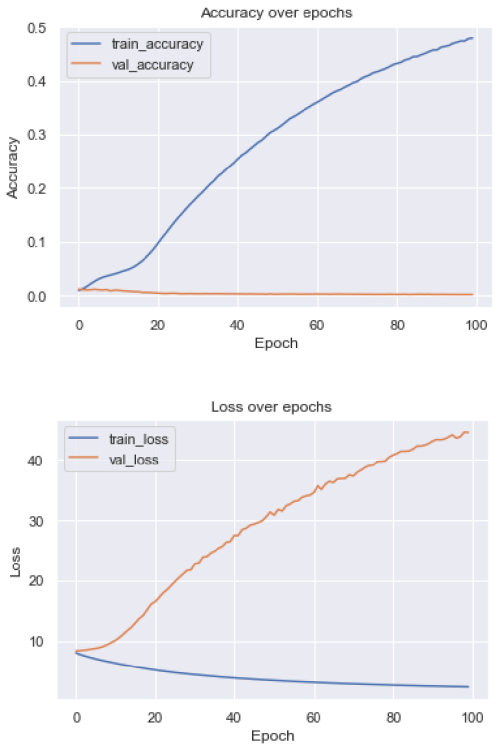

Figure 3: Train/Validation loss and accuracy on baseline model **(a)** accuracy **(b)** loss, as a function of epochs.

This is a good indicator of overfitting, as the model is unable to generalize on a validation set. In the end, the model had approximately 0.0001 validation accuracy. Considering there are ~ 12000 unique words in the data set, this baseline model barely performed above chance.

For our improved model (dropout, biLSTM), the training accuracy increased and the validation accuracy also increased early on until around epoch 20, but eventually plateaued and started decreasing as shown in Figure 4. Again, this is a sign of overfitting after epoch 20. However, unlike the baseline model, this new model achieved drastically higher validation accuracy at approximately 0.12. Although this might sound poor, 12% correct word prediction is quite impressive given the number of unique words in our data set. With a more complex (deeper) network, we would be able to achieve better results. But for the purposes of modeling children's word development, this model is sufficient.

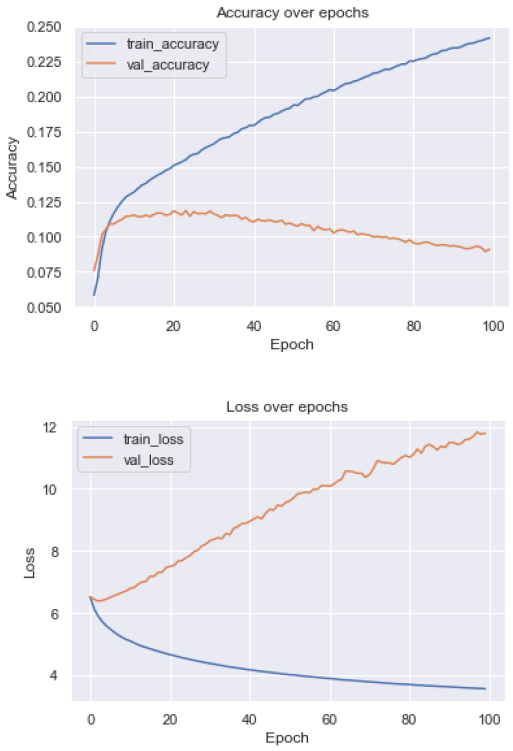

Figure 4: Train/Validation loss and accuracy on BiLSTM model **(a)** accuracy **(b)** loss, as a function of epochs.

Language Acquisition Modeling

To mimic the word production measure as provided by the CDI, we ran the models on all sequences and extracted the words that the model generated based on to obtain their vocabulary sizes.

The vocabulary size of the models showed rapid growth: the model that was trained for zero epochs only produced the world "little." The model that was trained for 20 epochs generated 580 unique words. When contrasted with norms of vocabulary size based on the CDI, this model's vocabulary size is comparable with that of the upper 50 percent of children 30 months of age. The model that was trained for 30 epochs generated 1569 unique words, which is comparable to 4-year old children (Owens, 2015). The model that was most extensively trained for 100 epochs had a vocabulary size of 8591 unique words, comparable to children in second and third grades, as there is variation in children (Segbers & Schroeder, 2017).

Next, we created semantic networks of these models. Semantic networks are graphs, which consist of a set of nodes and a set of edges that join pairs of nodes. An edge is an undirected link between two nodes and two nodes that are connected by an edge are neighbors. The degree of a node refers to the numbers of incoming and outgoing edges. We used cosine similarity to threshold the relationship between words in the graphs. Words were classified into nouns, adjectives, verbs, function words, and all other words, follow-

ing Wordbank. Figure 6 provides visual comparison of the semantic networks of the models trained for 10, 20, and 30 episodes and those of children of ages 18, 30, and 36 months, respectively, with a minimum cosine similarity of 0.4. Visual analysis of these networks showed interesting patterns. In the early semantic networks, similar concepts arise. For example, both children's and the LSTM's network exhibit a hub for the concept of family (such as "father" and "mommy"), and animals. Furthermore, utterances appear frequently in both networks, such as "baa" and "ha." The networks also exhibit hubs of densely connected words.

We further compared the average clustering coefficient between networks as a measure of network connectivity (see Eq. 1). The average clustering coefficient for the network of children 18 months of age was 0.35, similarly to the one of the LSTM model trained for 10 epochs with a clustering coefficient of 0.28. While the clustering coefficient remained the same in the network of children 30 months of age, it decreased to 0.23 in the model that was trained for 20 epochs. This reduction continued, so that the network of the model trained for 30 epochs showed a clustering coefficient of 0.22. Data for the children 36 months of age was not available, but visual analysis suggests that the degree of connectivity in children may indeed stronger than in the LSTM, as it exhibits several strong hubs.

$$C = \frac{1}{n} \sum_{v \in G} c_v \tag{1}$$

Equation 1: c = clustering coefficient, n = number of nodes in graph G, v = node.

To test whether early noun learning follows the principle of preferential attachment, we run a regression with degree at 100 epochs of training as the dependent variable and age of acquisition as the independent variable and found a significant relationship ($b = 2.42$, $t(147) = -0.18$, $p < 0.001$, see Figure 5). Following Hills et al. (2009), we plotted degree at 100 epochs relative to its cumulative distribution for comparison, showing the probability that a randomly chosen node was of degree equal to or higher than k. The plotted data showed a similar, but weaker power-law structure as in Hills et al. (2009). When fitting a power law to a data set, one should compare the goodness of fit to that of a log-normal distribution (Alstott, Bullmore, & Plenz, 2009). Our result suggests that a power-law distribution did not fit the data better than a log-normal distribution (log-likelihood ratio $= -17.75, p < 0001$), preventing the conclusion of similar underlying structures between the later semantic networks of the LSTM and children. Overall, while the semantic networks showed similar characteristics early in development, they diverged over the course of development. More specifically, while children's vocabulary exhibits power-law structures and indicates learning by preferential attachment, the evidence in the LSTM's networks is much weaker.

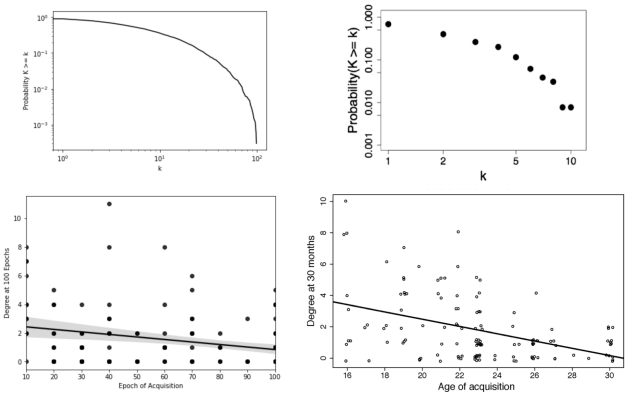

Figure 5: Upper plots: Log-log plot of the cumulative degree distribution for network after 100 epochs of training and the 30-month associative network in children (right). Lower plots: Degree of words at 100 epochs of training as a function of epoch of acquisition in the LSTM (left). Degree for each word in an associative network of 30-months year old children as a function of age of acquisition (left). Best-fitting regression lines are shown.

Discussion

Our results show that vocabulary growth in an LSTM unfolds very rapidly, reaching a vocabulary size of early schoolchildren after only 100 epochs of training. Despite these promising results, there are several limitations. First, our LSTM was not able to grasp different types of data. While it only learned from text data, children's language learning is supported by a variety of sources: conversations between people (auditory), illustrations in a picture book (visual), even experiences associated with words through touch (e.g., "pain" from touching a hot stove). Children that grow up in large households may experience more auditory words, and as a result have a different word development than children growing up in a small family. As such, environmental factors play an important role in language development.

More detailed analyses of early semantic networks showed similar characteristics in the early networks, but divergence with development. Children's semantic networks seem to grow by preferential attachment, while this pattern is not as clear in the networks of the LSTM. It can be argued that language learning by preferential attachment is not consistent across children's semantic networks and actual language acquisition follows different processes (Hills et al., 2009). Alternatively, it has been proposed that early vocabulary growth is driven by the connectivity of words in the learning environment to each other, instead of the connectivity of the words already in the vocabulary. If this model fit the data better, it would suggest that the LSTM learns those words that are central in the semantic and phonological environment and therefore, more noticeable. However, an important limiting factor is that the data in the network for comparison of growth re-

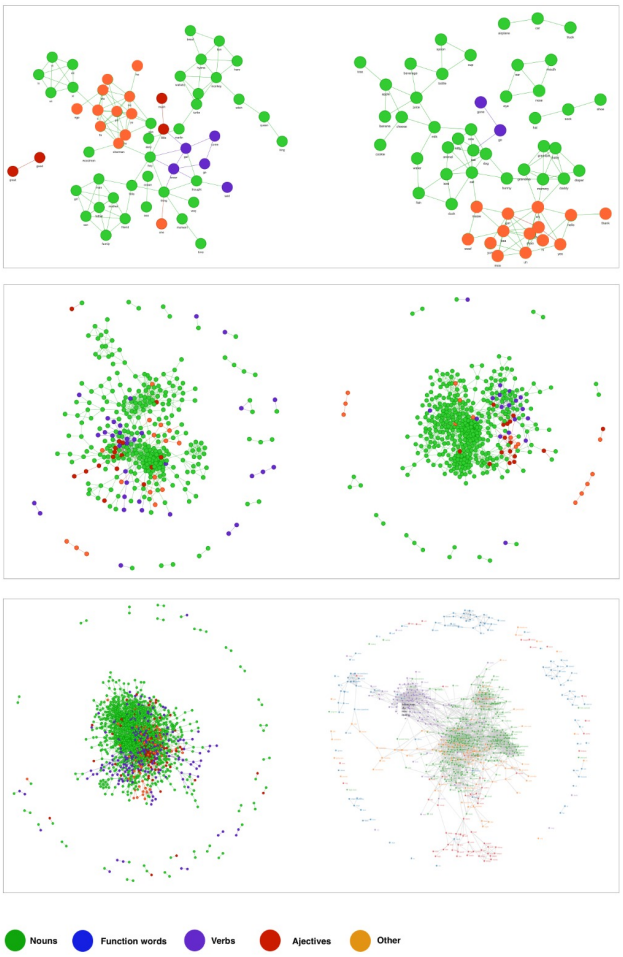

Figure 6: Semantic networks of the LSTM (left) and of children (right) with a minimum cosine similarity of 0.4. The upper graph shows the model after 10 epochs of training, and 18 months-old children respectively. The middle graph shows the model after 20 epochs of training, and 30 months-old children. The lower graph shows the model after 30 epochs of training, and 36 months-old children.

ported in Hills et al. (2009) was constructed from associative data and thus limits the validity of this comparison. We also did not test the model's language usage and understanding, which might elucidate further limitations in the LSTM. More recently, attention based models have replaced RNNs, but it is widely recognized that current neural networks are far from implementing human language abilities. Improving language models requires consideration of core human abilities, such as intuitive physics, intuitive psychology, and rapid learning with compositional causal models that help children acquire linguistic meaning and language from the very beginning (Lake, Ullman, Tenenbaum, & Gershman, 2017). Overall, in our LSTM models, semantic development seems to unfold differently compared to children, highlighting the different capabilities between human and current machine in the domain of language acquisition.

Code Availability

Code, original text files, and preprocessed data for this paper is available at https://github.com/jonathanlu02/lexical-classes-lstm.

References

Alstott, J., Bullmore, E., & Plenz, D. (2009). Powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv:1305.0215*.

Aylien News API. (2020). *Word embeddings and their challenges*. https://aylien.com/blog/word-embeddings-and-their-challenges. (Accessed: 2021-04-30)

Bloom, P. (2001). Précis of how children learn the meanings of words. *Behavioral and brain Sciences*, *24*(6), 1095.

Carey, S. (2009). The origin of concepts. *Oxford University Press.*.

Culurciello, E. (2018). *The fall of rnn/lstm.* https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0. (Accessed: 2021-05-04)

Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., & Reilly, J. (1993). The macarthur communicative development inventory: Words and sentences. *San Diego, CA: Singular*.

Fenson, L., et al. (2007). *Macarthur-bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, *20*(6), 729–739.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*(E253).

Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Top Cogn Sci*, *8*.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.

Mehler, J., Christophe, A., Ramus, F., Marantz, A., Miyashita, Y., & O'Neil, W. (2000). How infants acquire language: some preliminary observations. *Image, Language, Brain (Marantz, A. et al., eds)*, 51–75.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American*, *257*(3), 94–99.

Mujtaba, H. (2020). *What is word embedding — word2vec — glove.* https://www.mygreatlearning.com/blog/word-embedding/. (Accessed: 2021-05-04)

Owens, R. E. (2015). *Language development: An introduction (9th edition)* (9th ed.). Pearson.

Pastor-Satorras, R., Smith, E., & Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *Journal of Theoretical biology*, *222*(2), 199–210.

Rice, M. (1983). The role of television in language acquisition. *Developmental Review*, *3*(2), 211–224.

Segbers, J., & Schroeder, S. (2017). How many words do children know? a corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297–320.

Soni, M. (2020). *Understanding word embeddings from scratch — lstm model.* https://towardsdatascience.com/word-embeddings-and-the-chamber-of-secrets-lstm-gru-tf-keras-de3f5c21bf16. (Accessed: 2021-05-04)

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). A long short-term memory model for global rapid intensification prediction. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? a model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155.

Talmi, D., Lohnas, L. J., & Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological review*, *126*(4), 455.