MACS 30200
Delores Tang
Methods & Results

Methods

In this section, I would describe the data I used in this experiment and the machine learning tools I used as well as the neural network models I built for data analysis. This current study aims to extract features of words produced by children of different ages and with different language abilities. Transcripts would be analyzed in word level, and models in this study took words and part-of-speech (POS) tags as input and predicted either Language Impairment (Impaired vs. Normal) or Age (ranges from 5 years 0 months to 11 years 11 months).

Dataset

For this project, I used the Gillam corpus (Gillam & Pearson, 2004) from the open-sourced database, Childes (MacWhinney, 2000). Childes stands for the Child Language Data Exchange System, coming from a project started by MacWhinney and his colleagues in 1984 in order to establish a language acquisition database for researchers studying in various languages and all around the world.

Specifically, Gillam corpus contained linguistic data produced by 250 children with language impairment and 520 with normal language abilities. The original purpose of this corpus was to build the Test of Narrative Language (TNL; Gillam & Pearson, 2004), a test that aims to evaluate the quality of narratives produced by children and, therefore, to assist the diagnosis of language disorders. Linguistic data were obtained in a naturalistic study, in which children participants were asked to tell a story in a given scene (in McDonald’s in this experiment).

Children were also audiotaped during storytelling, and the audio files were transcribed later to CHAT format and uploaded to Childes under Gillam corpus.

Data preprocessing. In order to apply neural network models and other data analytical tools to the dataset, I preprocessed the data so that separate transcripts were combined into one file. First, I utilized childes-db (Sanchez et al., 2018), an API that helps to transform CHAT files from Childes database to more accessible and reproducible tabular format. I further processed the obtained tabular format to CSV format and kept only columns that are relevant to the current study: id (the identification number for each sentence), transcript_id (the identification number for each transcript/child), gloss (one gloss is a sentence produced by a child), part_of_speech (POS tags as defined by Childes), Impaired (whether a child has impaired = 1 or normal = 0 language abilities), Age, and Gender (female = 0 and male = 1).

After renaming and organizing columns of the dataset, each row represents one sentence produced by a child. Rows with empty gloss were dropped. As Gillam corpus dropped some children's data, only 668 children's data were analyzed. There are 497 children who had normal language abilities, and 171 had normal abilities; 323 were female and 345 were male. Children were ranged from age 5 to 11 years old, and specific statistics of age are included in Figure 1.

Preliminary Analysis

Before applying any deep learning models to the data, I first conducted some preliminary analysis using the frequencies of verbs and nouns children produced during storytelling as the input variable and their language abilities and Age as dependent variables. Logistic regression was used for the preliminary analysis. Besides verbs and nouns, frequencies of other POS tags were also reported in order to examine if there are patterns of other POS tags, besides verbs and nouns, that might be predicting features of the two output variables.

Word-Based Analysis

Preliminary analysis was conducted using different neural networks with the column "gloss" as input and either Language Impairment or Age as output.

Word embeddings. After processing the Gillam data, each word in gloss was vectorized to be able to use as inputs to complex machine learning models. Three embedding methods were used: GloVe, Word2Vec using a continuous Bag-of-Words (CBOW) architecture, and Word2Vec using a Skip-Gram (SG) architecture.

The Global Vectors for Word Representation (GloVe) is an unsupervised algorithm that represents words using pre-trained vectors (Pennington, Socher, & Manning, 2014). In this study, I used the word vector file "Glove.6B.zip", which was trained on Wikipedia 2014 and Gigaword to obtain the embedding matrix.

Word2Vec is another word vectorizing approach developed by Mikolov and colleagues (2013). Different from other word vectorizers, it produces vectors for a target word based on its context, specifically, words surrounding the target word. CBOW- and Skip-Gram-based Word2Vec are slightly different in how they define and predict the context of a given word. Two different embedding matrices were computed in this study for further word-based analysis.

Vocabulary. Each row in gloss was a sentence produced by a child, and it was regarded as one sample. In total, there are totally 24609 examples in the preprocessed data. The data was divided into a training set (68% of the data), a validation set (12% of the data), and a test set

(20% of data). A vocabulary of 3958 words was obtained from the dataset and was used to create three different embedding matrices.

POS-Tag-Based Analysis

Since the purpose of the study is to examine whether children of different ages learn certain words better and faster than other words, POS tags were more important than original words in the analysis.

POS bigrams. POS tags were not only analyzed in a single-tag manner, but they were also grouped into bigrams. Each bigram was indexed and vectorized separately using the TfidVectorizer provided by the Sklearn package in Python. This tool vectorizes each input, character, word, or bigram, according to its TF-IDF (Term Frequency-Inverse Document Frequency) score, taking into account the frequency of occurrence of a given input.

Model Construction

Naive Simple RNN model as a baseline. During analysis, I used a simple recurrent neural network model without pre-trained embedding matrices as a baseline model. Because I was interested in two output variables - Age and Language Impairment, two baseline models were trained and evaluated for both word- and POS tag-based analyses.

Two models are of the same architecture. Each model has one embedding layer, which transforms vocabulary index to word embeddings, as no pre-trained matrix was used. The model has three hidden layers, each with 64 neurons and an activation function of ReLU (Rectified Linear Unit). For Age, the output layer used Softmax function to perform a multinomial classification, while for Language Impairment, the output layer used a sigmoid function to perform binary classification.

LSTM-based RNN. Three bidirectional Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) models were then trained using three embedding matrices obtained respectively using GloVe, CBOW-based Word2Vec, and Skip-Gram-based Word2Vec for word-based analysis. For POS-tags, the same initial models were used despite that the input was vectorized tags and tag-bigrams.

By comparing the initial models, I selected one based on the reported validation loss and accuracy and built other LSTM-RNN models with modified numbers of hidden layers and neurons per layer. Because no overfitting was observed for word-based analysis, I did not use any dropout or regularizers.

Finally, the best model for each output variable was selected and evaluated on the held-out test set.

Results

Preliminary Results

(The data preprocessed for preliminary results had an error and thus did not match the data I used for other analyses. I would add on accurate statistics later).

Word-Based Results

The baseline model returned low maximum validation accuracy (*val_acc* = 0.24) and high minimum validation cost (*val_cost* = 2.23) in predicting Age, but it returned relatively better results for Language Impairment (*val_acc* = 0.78, *val_loss* = 0.50, see Figure 2).

Modified models returned similar results predicting Age, suggesting that either using single words produced by children participants might not give significant insights to what age group they belong. However, modified models using LSTM architecture predicting Language Impairment returned better results (See Table 1).

References

Gillam, R. B. & Pearson, N. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed Inc.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS.*

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2018, April 23). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System. Retrieved from psyarxiv.com/93mwx

Figure 1. Descriptive statistics of Age in Gillam corpus data.

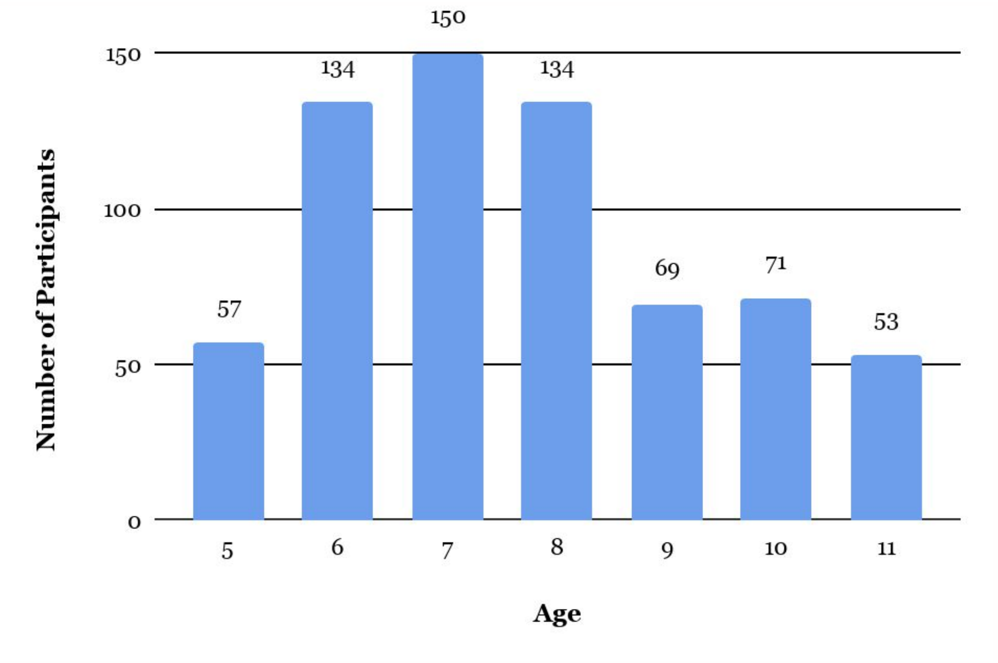

Figure 2. Validation error and loss across 20 epochs of training, reported by baseline model predicting Language Impairment.

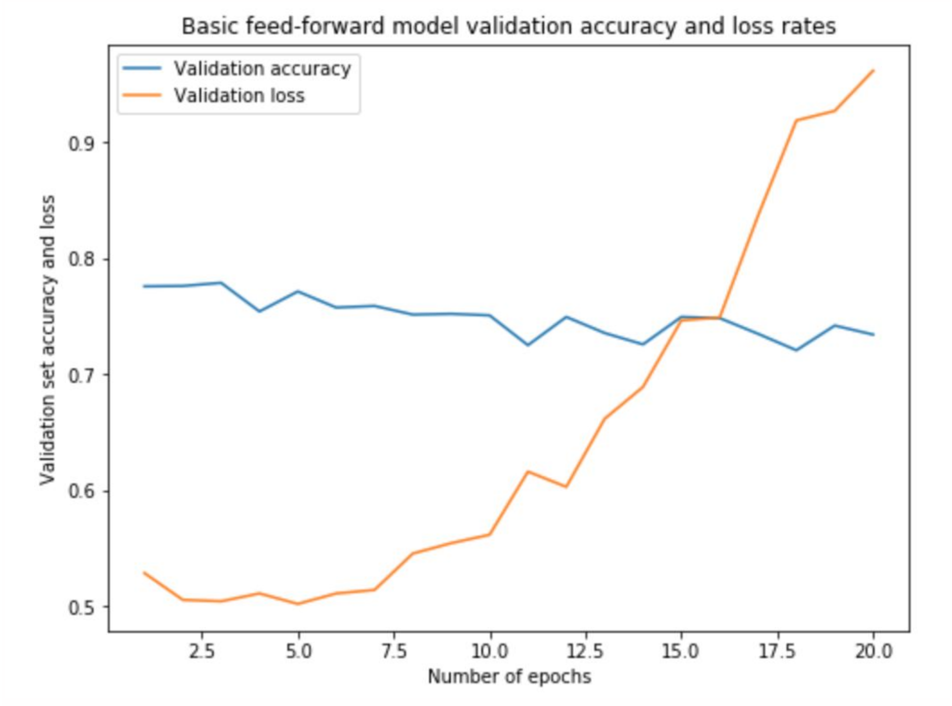

Table 1. Validation loss and accuracy using three models with LSTM architecture for predicting Language Impairment.

	Model 1 - GloVe	Model 2 - CBOW	Model 3 - SG
Parameters	LSTM architecture with 1 hidden layer Number of nodes per layer = 64 Number of epochs = 20 Batch size = 512		
Maximum validation accuracy	0.78	0.78	0.78
Minimum validation loss	0.61	0.66	0.63