Influence of input on part of speech tagging: Comparing adult directed and child directed speech

Niels Dickson
University of California, Irvine
nielsd[at]uci.edu

Abstract

The current work attempts to investigate the difference in syntactic structure between adult directed speech (ADS) and child directed speech (CDS). In particular, how might the learning from one speech type influence the syntactic predictions the learner makes when comprehending novel input. To begin this investigation into syntactic expectations, I created an unsupervised Hidden Markov Model part of speech tagger. I then trained one model on CDS from CHILDES corpora and another model on ADS from the Switchboard Corpus. Preliminary results suggest that, as expected, someone learning from CDS will find ADS samples less probable than CDS samples. Discussion of limitations, unsupervised POS tagging, and future work can be found in Conclusion.

1 Introduction

As we comprehend language, we make predictions about future input. In the framework of Information Theory, this tendency for humans to make predictions has been well documented for upcoming words and even predicting the next phoneme within a word. Evidence also suggests that humans make predictions about syntactic structure (Levy, 2008; Ferreira and Qiu, 2021), but this phenomenon is understudied in the Information Theory literature in comparison to semantic predictions. This disparity, no doubt, comes from the additional modeling difficulty in working with sentence structures as opposed to words. But strong evidence for humans engaging in structure prediction, and therefore making it a worthwhile topic of study, comes from the many studies investigating the garden path phenomenon. When processing a garden path sentence (e.g. "The horse raced past the barn fell"), we make a prediction about the structure of the sentence and the remaining words, but often in garden path sentences our predictions are incorrect producing the garden path phenomenon as we reparse the sentence. Additionally, studying this topic may have practical modeling advantages as some evidence suggests that separating the semantic and syntactic surprisal produces better performance in predicting reading time (Roark et al., 2009).

In the current work, I will be investigating predictions about syntactic structure with the tools from the Part of Speech (POS) tagging literature. Work in the POS tagging literature often frame the task as predicting the next POS given the current label, offering an interesting parallel to the topic of predicting syntactic structure. Specifically, I am interested in investigating what predictions about syntactic structure humans can learn from language input and how these predictions might change from adult-directed speech (ADS) to child-directed speech (CDS). Past syntactic analyses of CDS suggests that CDS may be less complex and more repetitive than ADS (for a review see Soderstrom 2007), but analyzing differences between CDS and ADS in terms of their expectations of syntactic categories can further our understanding of how CDS differs from ADS and the patterns that children can extract from their input. This research can also make predictions about differences in performance between children and adults in comprehension tasks.

2 Background

Part of Speech tagging is assigning a syntactic label to each word in some text. This process is traditionally a task of disambiguation where the model is given a dictionary to label most words deterministically but selects a label for some words that are syntactically ambiguous. Resolving the ambiguity is achieved using the context of the word.

A popular model for POS tagging is the Hidden Markov Model (HMM). A Markov Chain models

the probability of a sequence of observed objects using the simplifying assumption that the probability of the current object only depends on the previous object in the chain. An HMM is an extension of the Markov Model with the assumption that the sequence of observed states can be explained by some unobserved structure. Figure 1 illustrates how the observed states, w, are a result of the unobserved hidden states, y. We can see how this model naturally applies to the task of POS tagging where the observed states are words and the unobserved states are POS tags.

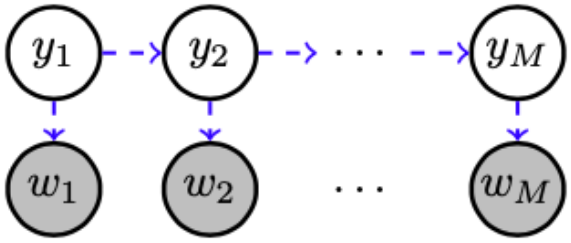

Figure 1: Graphical representation of an HMM taken from Eisenstein (2019).

The main components of this POS tagging implementation of HMM are the transition probabilities from label to label and the emission probabilities of a word given the label (Jurafsky and Martin, 2009). Given a tagged corpus, these probabilities can be calculated directly by counting. For the transition probabilities, one counts for the the relevant instances for a tag, t:

$$p(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

For the emission probabilities, one counts the relevant instances of a tag, t, and word, w:

$$p(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

We can see that this POS tagging method is highly supervised as we are using a corpus of gold standard labeling and a fixed set of POS tags. In investigating how humans process language, supervised learning is preferred. Additionally, unsupervised tagging has a practical modeling advantage in that the unsupervised model allows us to collapse POS labels into a standard set to use text from multiple corpora. The seminal work on unsupervised POS labeling comes from Goldwater and Griffiths (2007).

In Goldwater and Griffiths (2007), the authors formulate a Bayesian Trigram HMM POS tagger. In this model, the tags for a corpus of text is initialized at random, and the MAP tag sequence is found by iteratively resampling the tags to maximize the probability of the sequence (where the transition and emission probabilities follow a Dirichlet distribution). Much of their results report performance of the model run on fully tagged corpora (therefore giving the model dictionary knowledge), but the authors also manipulated the amount of dictionary knowledge provided. To do this, they collapsed the tags from the corpus into 17 categories and only gave the model access to the labeling of the most frequent words (where the frequency cut-off is being manipulated). Measuring accuracy for the fully unsupervised model (no dictionary knowledge) is difficult, but the model with the least dictionary information correctly tagged 49.7% of the words (random condition achieved 38.6% accuracy) illustrating the difficulty of unsupervised POS tagging.

3 Model

For the current project, I created an unsupervised POS tagging model trained on ADS and one trained on CDS. For samples of ADS, I used the Switchboard corpus (Godfrey et al., 1992) which complies thousands of phone conversations between adults. For samples of CDS, I used the CHILDES collection of corpora (MacWhinney, 2014) which compile speech that caregivers directed at their children. Specifically, I used the Valian corpus.

For the training, I extracted a selection of text (5,000 words) from each corpora and then divided it into 500 chunks with 10 words in each chunk. To create an HMM that could be tested on both CDS and ADS chunks, the vocabulary consisted of the unique words in both sets. Following Goldwater and Griffiths (2007), the states of my HMM consisted of 17 tags. I then randomly initialized, using a normal distribution sampler, the emission (length states X length vocabulary) and transition (length states X length of states) matrices. And to complete the HMM, I specified α to be the first row in the transition matrix (T) and ω to be the last row in T.

I then created two models from this HMM, one trained on ADS and one trained on CDS. To train the model on the relevant text, I wanted to adjust the emission (E) and transition (T) matrices to maximize the inside probability of the text. The inside probability is calculated by summing over all possible paths of our labels q:

$$p\left(\mathbf{w}_{t=1}^{K}\right) = \sum_{\mathbf{q} \in Q^{K+1}} p\left(\mathbf{q}, \mathbf{w}\right).$$

To calculate the inside probability, I used the following formula which includes a term, $f(q,t)$, that is calculated recursively:

$$p\left(\mathbf{w}_{t=1}^{K}\right) = \sum_{q \in Q} p_{\alpha}\left(q\right) f\left(q, 1\right)$$

$$f\left(q,t\right) = p_{E}\left(w_{t} \mid q\right) \sum_{q' \in Q} p_{T}\left(q' \mid q\right) f\left(q', t+1\right),$$

$$f\left(q, K+1\right) = p_{\omega}(q).$$

To maximize the inside probability for given text, I performed gradient descent on E and T using PyTorch. I could not run gradient descent on large text samples as the inside probability approached 0 as the length increased. So I decided to divide the text into chunks and iteratively apply gradient descent in which the output E and T matrices were fed forward to the application of gradient descent on the next text chunk. To achieve this, I decided to use fewer gradient steps (20) for each chunk but with a relatively large learning rate (0.05) in order to train on the 500 chunks. In piloting, I noticed that the model was achieving large inside probabilities by simply making E and T arbitrarily large, so the E and T were passed through a softmax function when calculating inside probability to constrain this behavior. Following this procedure, I trained one model on the ADS chunks and the other model on the CDS chunks.

As stated previously, estimating accuracy for unsupervised POS tagging is difficult because there is not an straightforward way to align the set of model tags with the set of gold standard tags. But we can analyze E and T to get a better understanding of the state of the model after training. First, for the ADS model, I wanted to see if there was any obvious clustering that could be extracted from the E output. To do this, I found the 50 most probable word emissions for each category. Using original tags from Switchboard, I found the percentage of these 50 words that were labelled as determiner, verb, or noun (the most common labels). This is plotted in figure 2.

Because these are the most common words in the input, it is reasonable to see that the three categories are represented in all the model categories

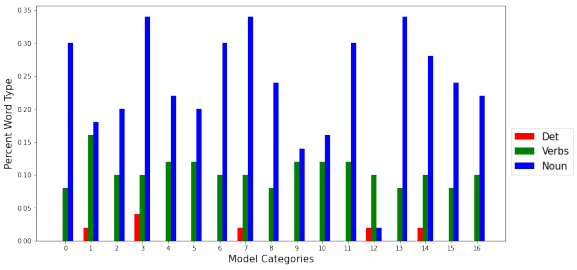

Figure 2: For each category in the ADS model, I found the 50 most probable word emissions. Of these 50, this plot illustrates the percent of words labeled as determiner, verb, or noun in the Switchboard corpus.

as opposed to seeing all nouns in one and all the determiners in another. More fine grained analyses of the less common tags may help clarify what kind of clustering this model prefers.

We can also visualize the T output to understand the transition preferences in the ADS model (seen in figure 3). Although these transitions are not reflective of transitions between parts of speech in human language, if the task is to maximize the inside probability, you might prefer to stay in one state that will eventually be your end state. Interestingly, this preferred state, 12, is the one that has a noticeably different distribution in figure 2. Maybe most of the legwork is done in state 12 such that the emissions from state 12 better reflect the observed distribution of the text. In the future, I could vary the number of categories to better understand this model's transition preferences, and I could try to force a preference for multiple states to better reflect human language.

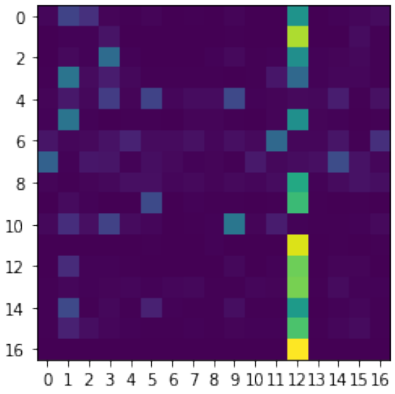

Figure 3: The probability of a transition from state on Y axis to state on X axis for the ADS model.

Now looking at the CDS model. I created similar visualizations, but used the Valian corpus labels for the category visualization.

We can again see similar performance in the T

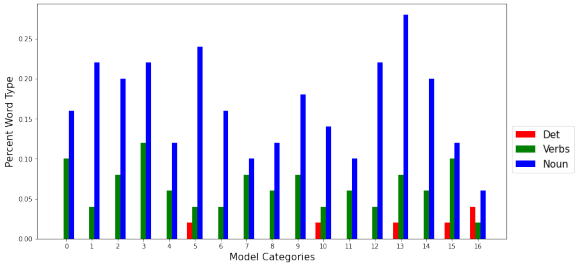

Figure 4: For each category in the CDS model, I found the 50 most probable word emissions. Of these 50, this plot illustrates the percent of words labeled as determiner, verb, or noun in the CHILDES corpus (Valian).

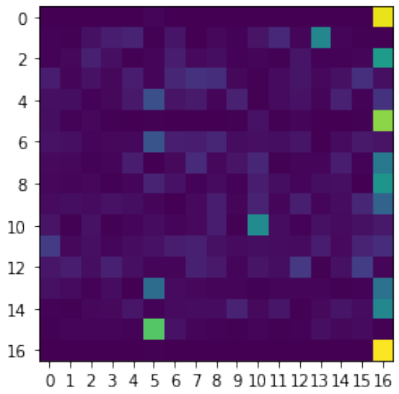

Figure 5: The probability of a transition from state on Y axis to state on X axis for the CDS model.

matrix as there is one state that is highly preferred (seen in figure 5). Looking at the categorization visual (4, we still do not see obvious clustering of our three categories, but we see the same interesting pattern of the highly preferred state having a noticeably different shape. It seems that all three common gold standard categories are underrepresented in the preferred model category suggesting again that this state better reflects the distribution of the text.

4 Results

Now returning to the main interest, we want to see how these two models perform on unseen text from the other text type. Taking another text sample (1,000 words) from each corpora, I created a set of unseen ADS text and a set of unseen CDS text (100 chunks of 10 words). The log inside probabilities are plotted in 6.

We can see that when the text from the opposite text type (ADS or CDS) from the model is judged as less probable than the text from the same text type. For example, the CDS model judges the CDS testing text to be more probable than the ADS testing text which was expected. Interestingly, there does not seem to be much of a difference between the two drop-offs from familiar text to unfamiliar text. Given the evidence that CDS is syntactically less complex, one might expect the ADS model to perform similarly on both text types.

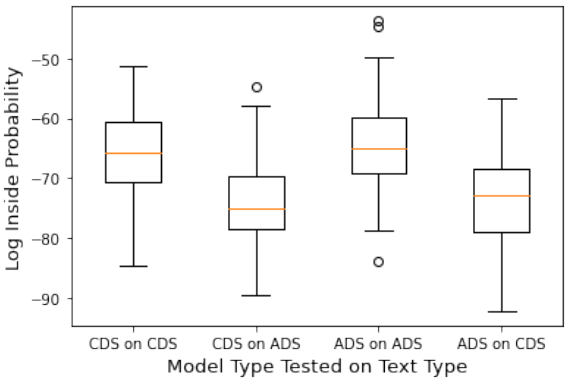

Figure 6: The log inside probability when CDS model tested on both text types (left) and when ADS model tested on both text types (right).

In interpreting these results, it is important to note that there are many factors relevant in comparisons across corpora. We see a noticeable difference in each model's performance on CDS versus ADS text, but this result could be a product of differences between corpora as opposed to differences between the structure of ADS and and the structure of CDS more generally. To increase confidence, I would need to compare across other ADS and CDS corpora.

5 Conclusion

To start, there are many limitations to the current study. First, the text trained on was relatively small (5,000 in current study compared to 24,000 in Goldwater and Griffiths 2007). Second, in order to achieve the goal of better understanding the structure in ADS and CDS that might lead to different syntactic predictions, I should have accompanied my model with one in the literature (i.e. the model created by Goldwater and Griffiths 2007). I don't have a great reason for not implementing the Goldwater-Griffiths (GG) model other than time constraints. But it is still informative to reflect on the differences between the unsupervised approaches. One obvious downside to my model in comparison to the GG model is that it isn't maximizing the string of tags for a large body of text. It seems that accurate POS tagging will require some sampling as seen in the GG model Gibbs

sampler where the tags are iteratively resampled to find the MAP tag sequence of the whole text. One potential advantage of my model is that it incrementally takes in input and adjusts the E and T matrices which may mirror the incrementally of language processing, but I would need to perform more analyses to determine how drastically the E and T matrices are tuned for each text sample. One way to maximize the inside probability on a more global scale using my model is to train on small text windows and slowly increase the size.

Despite these limitations, the current work further highlights the difficulties in unsupervised POS tagging in terms of implementation and estimating accuracy. It also, rather fortuitously, contributes a model that incrementally learns about the hidden labels influencing text sequences. From the preliminary results reported here, we expect to see differences between how children and adults form predictions about upcoming syntactic categories. Framing the phenomenon of predicting syntactic structure in the context of POS tagging, allows us to make predictions about someone's expectation of the following syntactic structure which may contribute to reading times in language comprehension tasks. To extend this modelling work, future work can test a wider range of corpora and POS tagging models to better understand how humans form syntactic predictions from their input.

References

Jacob Eisenstein. 2019. *Introduction to natural language processing*. MIT press.

Fernanda Ferreira and Zhuang Qiu. 2021. Predicting syntactic structure. *Brain Research*, page 147632.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic. Association for Computational Linguistics.

D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 324–333.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.