An incremental model of syntactic bootstrapping

Christos Christodoulopoulos[*], **Dan Roth**[*] and **Cynthia Fisher**[†]
[*]Department of Computer Science [†]Department of Psychology
University of Illinois at Urbana-Champaign
{christod,danr,clfishe}@illinois.edu

Abstract

Syntactic bootstrapping is the hypothesis that learners can use the preliminary syntactic structure of a sentence to identify and characterise the meanings of novel verbs. Previous work has shown that syntactic bootstrapping can begin using only a few seed nouns (Connor et al., 2010; Connor et al., 2012). Here, we relax their key assumption: rather than training the model over the entire corpus at once (*batch mode*), we train the model incrementally, thus more realistically simulating a human learner. We also improve on the verb prediction method by incorporating the assumption that verb assignments are stable over time. We show that, given a high enough number of seed nouns (around 30), an incremental model achieves similar performance to the batch model. We also find that the number of seed nouns shown to be sufficient in the previous work is not sufficient under the more realistic incremental model. The results demonstrate that adopting more realistic assumptions about the early stages of language acquisition can provide new insights without undermining performance.

1 Introduction

An important aspect of how children acquire language is how they map lexical units and their combinations to underlying semantic representations (Gleitman, 1990). Syntactic bootstrapping is an account of this aspect of language learning. It is the hypothesis that learners can use the syntactic structure of a sentence to characterise the meanings of novel verbs. However, the problem remains of how learners first identify verbs, and characterise the syntactic structure of sentences.

One mechanism for resolving this issue is Structure Mapping (Fisher et al., 2010), which hypothesises that, assuming an innate one-to-one mapping between nouns and semantic arguments in an utterance, children are able to use this information to first identify verbs and their arguments, and then assign semantic roles to those arguments. In this paper we provide a computational model for this account of syntactic bootstrapping. We use a system called *BabySRL* (Connor et al., 2010; Connor et al., 2012) that assigns semantic roles to arguments in an utterance – a simplified version of the Semantic Role Labeling Task (SRL; (Palmer et al., 2011)). Here, we focus on the preliminary task of identifying nouns and verbs from sentences in a corpus of child-directed speech (the Brown corpus (Brown, 1973), a subset of the CHILDES database (MacWhinney, 2000)). Previous work (Connor et al., 2010) presented a model which could identifying noun and verb clusters with minimal supervision (a few seed nouns). However, this model had two substantial limitations: the first was training was done in a *batch* mode, where the entire dataset was made available to the learner before any predictions were made; the second was that while the noun prediction was aggregated (previously identified known clusters persisted throughout the run through the data), the verb prediction was not (previously identified verb clusters had no effect on future predictions).

The current work makes two main advances on the previous work. Firstly, it addresses the batch mode limitation, adopting a more cognitively plausible approach where all sentences are given to the learner incrementally, more accurately modelling ongoing learning from child-directed speech. Secondly, it adopts an aggregated approach to verb prediction, as described in section 2.2, which capitalises on the fundamental assump-

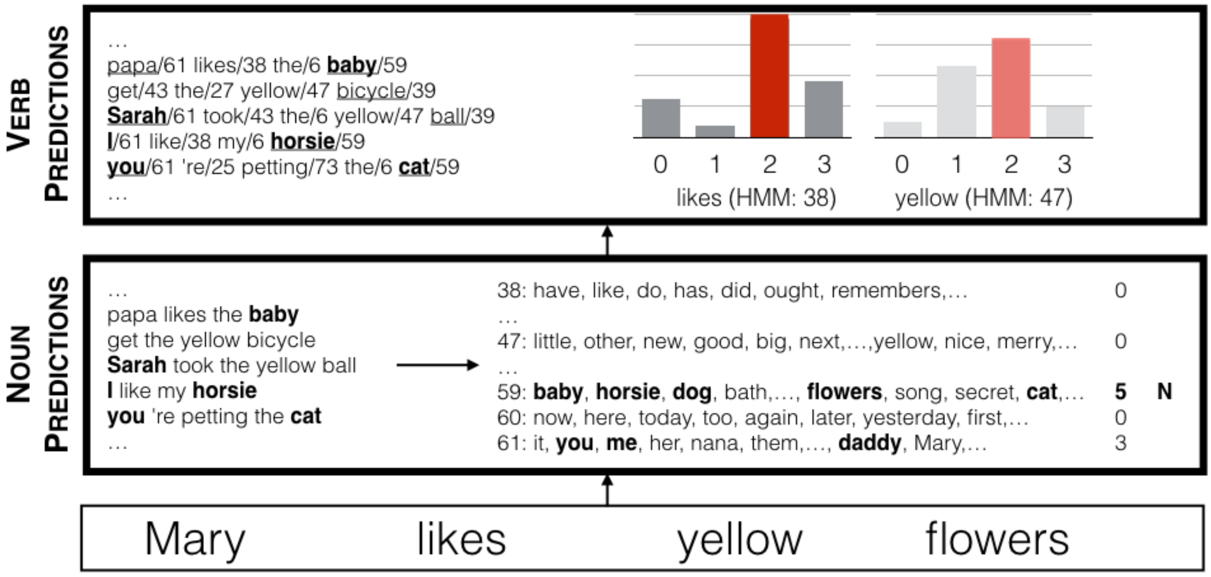

VERB PREDICTIONS
...
papa/61 likes/38 the/6 baby/59
get/43 the/27 yellow/47 bicycle/39
Sarah/61 took/43 the/6 yellow/47 ball/39
I/61 like/38 my/6 horsie/59
you/61 're/25 petting/73 the/6 cat/59
...
0 1 2 3
likes (HMM: 38)
0 1 2 3
yellow (HMM: 47)
NOUN PREDICTIONS
...
papa likes the baby
get the yellow bicycle
Sarah took the yellow ball
I like my horsie
you 're petting the cat
...
38: have, like, do, has, did, ought, remembers,... 0
...
47: little, other, new, good, big, next,...,yellow, nice, merry,... 0
...
59: baby, horsie, dog, bath,..., flowers, song, secret, cat,... 5 N
60: now, here, today, too, again, later, yesterday, first,... 0
61: it, you, me, her, nana, them,..., daddy, Mary,... 3
Mary likes yellow flowers

Figure 1: Illustration of the noun and verb prediction heuristics. The noun heuristic stage receives words assigned to HMM states and a list of seed nouns, and assigns the noun label to states that contain 4 or more seed nouns, assumed to be learned without syntactic help (right-hand side columns show the number of identified seeds and assignment). The verb heuristic receives a list of noun states per sentence and accumulates counts of co-occurring nouns for each of the non-noun states (right-hand side histograms). It assigns the verb label to the state with the highest probability of occurring with the number of nouns that appear in the sentence.

tion that distributional clusters will behave in a grammatically consistent fashion (“once a verb, always a verb”).

2 Noun and verb prediction

Figure 1 describes the heuristics for noun and verb prediction. Firstly, we model the distributional-based word categorization with a hidden Markov model (HMM) using 80 states. We used a Variational Bayes HMM model (Beal, 2003), trained off-line over a very large corpus of child-directed speech (2.1M tokens). We then use the method described in Connor et al. (2010) to identify which of these HMM states act as arguments (nouns) and predicates (verbs). As in the original work, we also give the HMM a number of function words as identified by their part-of-speech tags in order to be clustered into separate reserved states. This represents (but does not model explicitly) the assumption that infant learners can identify function words based on a variety of cues, including linguistic context, prosody, and frequency (Gerken and McIntosh, 1993; Christophe et al., 2008; Hochmann, 2013). Note also, that the list of function words was given to the HMM during training and not during the tagging of the BabySRL corpus. This means that for this corpus, the HMM is using the same distributional statistics as for the content words to decide on the function-word state membership.

2.1 Identifying nouns

As in Connor et al. (2010), we use a simple heuristic to identify noun HMM states. We assume a number of (up to 75) “seed” nouns (taken from Dale and Fenson (1996) – we chose the words that were produced by at least 50% of children under 21 months old). These words, assumed to be learned without syntactic knowledge, are recognised by the learner as verb arguments by virtue of structure-mappings one-to-one mapping assumption (Fisher et al., 2010). Using that knowledge, the learner is able to identify which HMM states contain these nouns and label them as arguments. Any state that contains 4 or more seed nouns is labelled as a noun state. We also experimented with a *dynamic* noun threshold: rather than keeping it to a fixed number (4), we used a number of functions that would dynamically increase this threshold according to the number of seed nouns presented to the learner. Experiments that increased the threshold up to 30 with linear, exponential, or logarithmic functions revealed no significant difference in results.

2.2 Identifying verbs

After running the noun heuristic, each remaining word (that does not belong to a function-word HMM state) is considered a candidate verb. For the purposes of this process, we assume that there is a single verb for each utterance. However, we use all the sentences available in the BabySRL corpus, a bare majority of which (51%) have only one verb predicate.

For the verb identification heuristic, we create a histogram of the number of times each non-noun content word (verb candidate) co-occurs with a specific number of noun arguments (shown in the top right of Figure 1). After this stage, as discussed in the Introduction, we diverge from the original model and adopt an aggregated prediction policy. The original model simply chose the "winner" of the histogram-based predictions: the candidate i with state s_i that maximized the probability of the identified number of noun arguments. For this new model, instead of assigning the verb label directly to the winner, we aggregate the predictions for each sentence into two numbers: the number of times state s_i was chosen as the the winner of the histogram-based predictions $(\#s_i(pred))$, and the overall number of times state s_i appeared in the corpus $(\#s_i(\cdot))$. From these two numbers we can calculate the probability of this state being a "stable" verb, $p(s_i(pred)) = (\#s_i(pred)/\#s_i(\cdot))$. For each sentence, we then pick the candidate whose state has the highest probability of being a stable verb. If multiple candidates have the same state and therefore the same probability, we choose the first.[1]

One of the corollaries of this experiment is that for the verb heuristic to work, the true argument structure of a verb (number of *core* arguments) has to align with the number of predicted arguments (nouns). To verify this, we looked at the number of times a verb's core arguments agree with the number of gold-standard nouns. We found that this is true for 36.3% of the sentences with a single verb (30.6% overall). This seemingly low score reflects the fact that not all arguments are single nouns: some contain no nouns, (as in the adjective argument of "looks nice"), and some contain multiple nouns, mostly in the form of conjunctions ("the boy and the girl").[2] The implication here is that if the verb heuristic was only using the count histograms as a source of information, its performance would have been mediocre. However, by excluding noun and function words states as potential arguments, the verb heuristic is able to achieve a pretty robust precision as we will see in section 4.

3 Incremental prediction

During language acquisition, children are exposed to learning data incrementally, meaning they are not exposed to all the data before having to generate their own hypotheses. To model this incremental exposure, the following changes to the original model were made.

Rather than noun prediction preceding verb prediction, in the incremental model both processes happen concurrently. When the model is exposed to the first sentence, it will identify no noun states because none of them exceed the threshold of 4 seed nouns. However, if a seed noun occurs, its appearance will be counted towards the sum of its state.

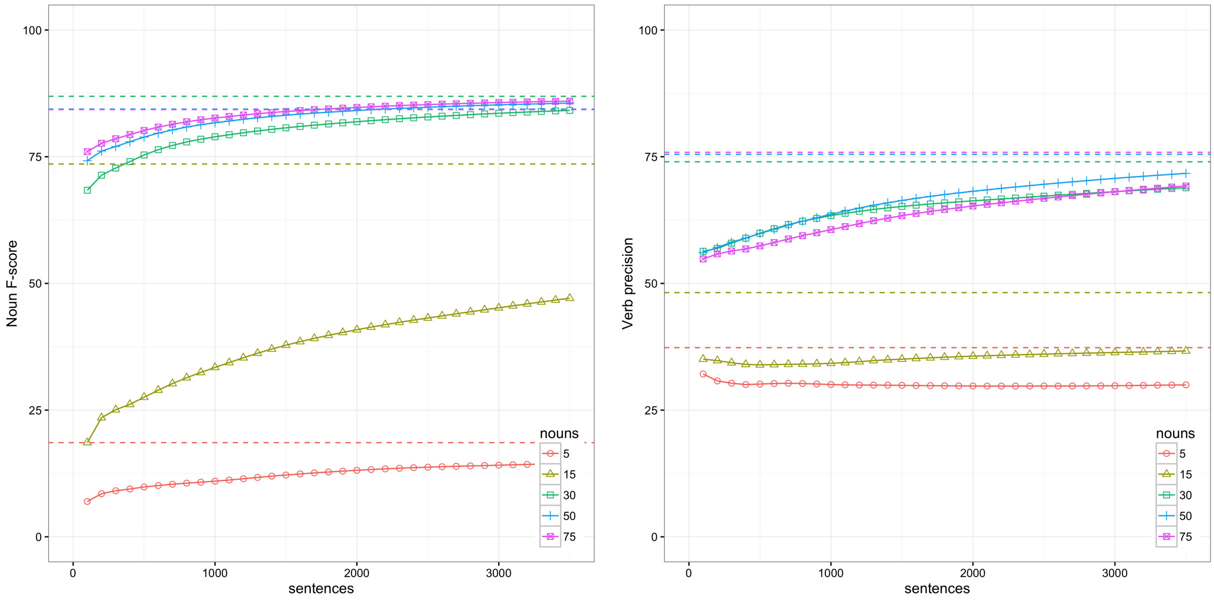
For example, in a case where the first four utterances in the corpus are as follows (HMM states are indicated by numbers following their corresponding word, function-word states are in grey and seed nouns are in bold):

(1) a. papa/57 wants/58 an/6 **apple**/39
b. get/43 the/27 red/79 **bicycle**/39
c. come/75 and/21 move/43 **horsie**/39
d. **i**/50 forgot/63 a/6 **spoon**/39
e. **you**/50 're/25 eating/73 the/6 broom/39

When the model reaches utterance (1-a), it recognises the seed noun 'apple', and so increments the counter for state 39. The only information available to the verb prediction module at this point is that 'apple', as a seed noun, is a potential noun. Therefore, this sentence contains two possible verbs,'papa' and 'wants' ('a/an' has a known function-word state). Therefore, both states 57 and 58 are stored in the verb histograms as having one argument and since it appears first (see foot-

[1] This method could allow us to predict multiple verbs per sentence, if instead of assigning the state with the highest probability, we set a threshold over which every state is assigned the verb label.

[2] Compound nouns ("ice cream" or "fire truck") are discounted using a simple heuristic of joining contiguous noun mentions.



(a) Noun F-score

(b) Verb precision

Figure 2: Results from the incremental noun/verb prediction, averaged over three children from the Brown corpus (solid lines). The x axis shows the number of sentences. Colours indicate number of seed nouns. For reference, dotted lines show results from the batch mode heuristics over all sentences including those with multiple verbs, using the same verb aggregation techniques described in section 2.2. For nouns, F-score is used, since the model predicts multiple nouns per utterance. For verbs, since only one verb is predicted per utterance, precision is used as the evaluation metric.

note 1), 'papa' will be chosen as the verb.[3]

The process repeats for utterances (1-b)–(1-d), each of which contains one seed noun in state 39. When the system reaches utterance (1-e), state 39 will have attained the threshold of 4 identified nouns. Utterance (1-e) therefore contains one noun identified via this noun heuristic, 'broom', and one known seed noun, 'you', leaving 'eating' as the only allowable verb candidate, and correctly predicting the argument histogram count (2) for its state (73). Using this toy example, we can see how it will not take long for both the noun and verb heuristics to reach the prediction level of the batch mode via an incremental process.

Note that while noun and verb prediction is truly incremental, the preliminary HMM learning and state assignments happen in batch mode. This assumption could be relaxed in future, since there already exist incremental models of word category assignment (Parisien et al., 2008; Fountain and Lapata, 2011). Here, as with the original work, we chose not to focus on this earlier stage of language acquisition, and instead assume that learning distributional facts about words proceeds largely independently for some time, until a few nouns are known – at which point syntax guides interpretation of the distributional classes. However, we know that category learning itself is influenced by syntactic properties (Christodoulopoulos et al., 2012). As such, in future work we plan to integrate the syntactic category learning with the verb and noun prediction stage to improve the accuracy of both.

4 Results and Discussion

We now present the results of the two main advances over the previous work of Connor et al. (2010): the incremental version of the verb and noun heuristics, and the aggregated predictions for the verb heuristic.

Figure 2 shows the results from the two tasks of noun and verb prediction averaged over three chil-

[3] The storing of both states 57 and 58 as potential one-argument verbs in the example may seem to conflict with the assumption that there is only one verb per sentence. It is true that at this stage, the model will lose information relevant to the true number of arguments of each verb, since potential arguments may be wrongly identified as verb candidates. However, the statistical stability of verb argument-taking behaviour, as well as the incrementally improving noun heuristic, leads to these early errors being corrected. In addition, this approach leaves space for a future version of this model where multiple verbs per sentence can be predicted.

dren, as well as the results of the original batch version from Connor et al. (2010). It is worth noting that the three children in the Brown corpus had different numbers of sentences that came from different age ranges. As such, the average trajectories mask substantial individual differences. There are two main findings: 1) the incremental scores for each number of seed nouns slowly converge to those of the batch mode; 2) similar to the original study, there is a plateau for both noun and verb prediction scores around 30 seed nouns.

For the noun prediction, we can see that the number of seed nouns it takes to reach comparable performance is slightly higher than in the batch mode model. For instance, with 15 seed nouns, the incremental prediction achieves a score of 47.1%, whereas the batch mode achieves a score of 73.6%. This is important, because it shows that the number of seed nouns the batch mode suggested was sufficient is not sufficient under a more realistic incremental model. Interestingly, this difference is not as pronounced for the verb prediction scores. The reason for this is that by aggregating over the histogram-based predictions, we can recover from more noise coming from the noun assignment. We also replicated the original (non-aggregated) verb heuristic from Connor et al. (2010). The results follow similar trends, although the absolute numbers are lower. This is verified our intuition that the grammatical 'meaning' of HMM states is indeed stable.

This work also raises a more general point about computational models of language learning. Real human learners not only have limited resources such as memory and processing power, but also are exposed to training instances incrementally and only once. Related work in the field of computer vision tries to mimic these learning conditions ("one-shot learning", Fei-Fei et al. (2006)), but this approach has not yet attracted much attention in the field of computational modeling of language acquisition.[4] We present these results as a preliminary step in this direction, showing that we can still attain good performance even while acknowledging these limitations, and that this can give us more insights into what exactly human learners require to support acquisition.

5 Conclusion

In this paper, we presented an incremental version of the syntactic bootstrapping model of Connor et al. (2010), with the additional innovation of aggregating over verb predictions – the latter representing the fundamental assumption that the tagging of HMM states with grammatical category "meaning" is stable ("once a verb, always a verb"). We showed that given a high enough number of seed nouns, an incremental model can achieve similar performance within around 2000 sentences for noun predictions and 3000 sentences for verb predictions. Importantly, the results also show that the number of seed nouns shown to be sufficient in the previous work is not sufficient under a more realistic model where the learner encounters data incrementally. More broadly, we demonstrate that adopting more realistic assumptions about the early stages of language acquisition can tell us more about what learners require to bootstrap the acquisition of syntactic categories while maintaining high performance.

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions. Many thanks also to Catriona Silvey for her help with the manuscript. This research is supported by NIH grant R01-HD054448-07.

References

Matthew James Beal. 2003. *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University of London.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2012. Turning the pipeline into a loop: Iterated unsupervised dependency parsing and pos induction. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 96–99, June.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and speech*, 51(1-2):61–75.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. 2010. Starting from Scratch in Semantic Role Labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, jul. Association for Computational Linguistics.

[4] A notable exception is the work on incremental word category acquisition mentioned above (Parisien et al., 2008; Fountain and Lapata, 2011).

Michael Connor, Cynthia Fisher, and Dan Roth. 2012. Starting from scratch in semantic role labeling: Early indirect supervision. In A. Alishahi, T. Poibeau, and A. Korhonen, editors, *Cognitive Aspects of Computational Language Acquisition*. Springer.

Philip S. Dale and Larry Fenson. 1996. Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28(1):125–127.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611.

Cynthia Fisher, Yael Gertner, Rose M Scott, and Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149, mar.

Trevor Fountain and Mirella Lapata. 2011. Incremental models of natural language category acquisition. In *Proceedings of the 32st Annual Conference of the Cognitive Science Society*.

LouAnn Gerken and Bonnie J McIntosh. 1993. Interplay of function morphemes and prosody in early language. *Developmental psychology*, 29(3):448.

Lila Gleitman. 1990. The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55, jan.

Jean-Rémy Hochmann. 2013. Word frequency, function words and the second gavagai problem. *Cognition*, 128(1):13–25, jul.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2011. *Semantic Role Labeling*. Morgan & Claypool Publishers, feb.

Christopher Parisien, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental bayesian model for learning syntactic categories. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 89–96. Association for Computational Linguistics.