An Incremental Bayesian Model for Learning Syntactic Categories

Christopher Parisien, Afsaneh Fazly and Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, ON, Canada
[chris,afsaneh,suzanne]@cs.toronto.edu

To appear in Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008)

Abstract

We present an incremental Bayesian model for the unsupervised learning of syntactic categories from raw text. The model draws information from the distributional cues of words within an utterance, while explicitly bootstrapping its development on its own partially-learned knowledge of syntactic categories. Testing our model on actual child-directed data, we demonstrate that it is robust to noise, learns reasonable categories, manages lexical ambiguity, and in general shows learning behaviours similar to those observed in children.

1 Introduction

An important open problem in cognitive science and artificial intelligence is how children successfully learn their native language despite the lack of explicit training. A key challenge in the early stages of language acquisition is to learn the notion of abstract syntactic categories (e.g., nouns, verbs, or determiners), which is necessary for acquiring the syntactic structure of language. Indeed, children as young as two years old show evidence of having acquired a good knowledge of some of these abstract categories (Olguin and Tomasello, 1993); by around six years of age, they have learned almost all syntactic categories (Kemp et al., 2005). Computational models help to elucidate the kinds of learning mechanisms that may be capable of achieving this feat. Such studies shed light on the possible cognitive mechanisms at work in human language acquisition, and also on potential means for unsupervised learning of complex linguistic knowledge in a computational system.

Learning the syntactic categories of words has been suggested to be based on the morphological and phonological properties of individual words, as well as on the distributional information about the contexts in which they appear. Several computational models have been proposed that draw on one or more of the above-mentioned properties in order to group words into discrete unlabeled categories. Most existing models only intend to show the relevance of such properties to the acquisition of adult-like syntactic categories such as nouns and verbs; hence, they do not necessarily incorporate the types of learning mechanisms used by children (Schütze, 1993; Redington et al., 1998; Clark, 2000; Mintz, 2003; Onnis and Christiansen, 2005). For example, in contrast to the above models, children acquire their knowledge of syntactic categories incrementally, processing the utterances they hear one at a time. Moreover, children appear to be sensitive to the fact that syntactic categories are partially defined in terms of other categories, e.g., nouns tend to follow determiners, and can be modified by adjectives.

We thus argue that a computational model should be incremental, and should use more abstract category knowledge to help better identify syntactic categories. Incremental processing also allows a model to incorporate its partially-learned knowledge of categories, letting the model *bootstrap* its development. To our knowledge, the only incremental model of category acquisition that also incorporates bootstrapping is that of Cartwright and Brent (1997). Their template-based model, however, draws on very specific linguistic constraints and rules to learn categories. Moreover, their model has difficulty with the variability of natural language data.

We address these shortcomings by developing an incremental probabilistic model of syntactic category acquisition that uses a domain-general learning algorithm. The model also incorporates a bootstrapping mechanism, and learns syntactic categories by looking only at the general patterns of distributional similarity in the input. Experiments performed on actual (noisy) child-directed data show that an explicit bootstrapping component improves the model's ability to

learn adult-like categories. The model's learning trajectory resembles some relevant behaviours seen in children, and we also show that the categories that our model learns can be successfully used in a lexical disambiguation task.

2 Overview of the Computational Model

We adapt a probabilistic incremental model of unsupervised categorization (i.e., clustering) proposed by Anderson (1991). The original model has been used to simulate human categorization in a variety of domains, including the acquisition of verb argument structure (Alishahi and Stevenson, 2008). Our adaptation of the model incorporates an explicit bootstrapping mechanism and a periodic merge of clusters, both facilitating generalization over input data. Here, we explain the input to our model (Section 2.1), the categorization model itself (Section 2.2), how we estimate probabilities to facilitate bootstrapping (Section 2.3), and our approach for merging similar clusters (Section 2.4).

2.1 Input Frames

We aim to learn categories of words, and we do this by looking for groups of similar word usages. Thus, rather than categorizing a word alone, we categorize a word token *with* its context from that usage. The initial input to our model is a sequence of unannotated utterances, that is, words separated by spaces. Before being categorized by the model, each word usage in the input is processed to produce a *frame* that contains the word itself (the head word of the frame) and its distributional context (the two words before and after it). For example, in the utterance 'I gave Josie a present,' when processing the head word *Josie*, we create the following frame for input to the categorization system:

feature	w_{-2}	w_{-1}	w_0	w_{+1}	w_{+2}
	I	gave	**Josie**	a	present

where w_0 denotes the head word feature, and w_{-2}, w_{-1}, w_{+1}, w_{+2} are the context word features. A context word may be 'null' if there are fewer than two preceding or following words in the utterance.

2.2 Categorization

Using Anderson's (1991) incremental Bayesian categorization algorithm, we learn clusters of word usages (i.e., the input frames) by drawing on the overall similarity of their features (here, the head word and the context words). The clusters themselves are not predefined, but emerge from similarities in the input. More formally, for each successive frame F in the input, processed in the order of the input words, we place F into the most likely cluster, either from the K existing clusters, or a new one:

$$\text{BestCluster}(F) = \underset{k}{\text{argmax}}\, P(k|F) \qquad (1)$$

where $k = 0, 1, .., K$, including the new cluster $k = 0$. Using Bayes' rule, and dropping $P(F)$ from the denominator, which is constant for all k, we find:

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k) \qquad (2)$$

The prior probability of k, $P(k)$, is given by:

$$P(k) = \frac{cn_k}{(1-c)+cn}, \qquad 1 \leq k \leq K \quad (3)$$

$$P(0) = \frac{1-c}{(1-c)+cn} \qquad (4)$$

where n_k is the number of frames in k, and n is the total number of frames observed at the time of processing frame F. Intuitively, a well-entrenched (large) cluster should be a more likely candidate for categorization than a small one. We reserve a small probability for creating a new cluster (Eq. 4). As the model processes more input overall, it should become less necessary to create new clusters to fit the data, so $P(0)$ decreases with large n. In our experiments, we set c to a large value, 0.95, to further increase the likelihood of using existing clusters.[1]

The probability of a frame F given a cluster k, $P(F|k)$, depends on the probabilities of the features in F given k. We assume that the individual features in a frame are conditionally independent given k, hence:

$$P(F|k) = P_H(w_0|k) \prod_{i \in \{-2,-1,+1,+2\}} P(w_i|k) \quad (5)$$

where P_H is the head word probability, i.e., the likelihood of seeing w_0 as a head word among the frames in cluster k. The context word probability $P(w_i|k)$ is the likelihood of seeing w_i in the i^{th} context position of the frames in cluster k. Next, we explain how we estimate each of these probabilities from the input.

2.3 Probabilities and Bootstrapping

For the head word probability $P_H(w_0|k)$, we use a smoothed maximum likelihood estimate (i.e., the proportion of frames in cluster k with head word w_0). For the context word probability $P(w_i|k)$, we can form two estimates. The first is a simple maximum likelihood estimate, which enforces a preference for creating clusters of frames with the same context words. That is, head words in the same cluster will

[1] The prior $P(k)$ is equivalent to the prior in a Dirichlet process mixture model (Sanborn et al., 2006), commonly used for sampling clusters of objects.

tend to share the same adjacent words. We call this word-based estimate P_{word}.

Alternatively, we may consider the likelihood of seeing not just the context word w_i, but *similar* words in that position. For example, if w_i can be used as a noun or a verb, then we want the likelihood of seeing *other* nouns or verbs in position i of frames in cluster k. Here, we use the partial knowledge of the learned clusters. That is, we look over all existing clusters k', estimate the probability that w_i is the head word of frames in k', then estimate the probability of using the head words from those other clusters in position i in cluster k. We refer to this category-based estimate as P_{cat}:

$$P_{cat}(w_i|k) = \sum_{k'} P_H(w_i|k')P_i(k'|k) \qquad (6)$$

where $P_i(k'|k)$ is the probability of finding usages from cluster k' in position i given cluster k. To support this we record the categorization decisions the model has made. When we categorize the frames of an utterance, we get a sequence of clusters for that utterance, which gives additional information to supplement the frame. We use this information to estimate $P_i(k'|k)$ for future categorizations, again using a smoothed maximum likelihood formula.

In contrast to the P_{word} estimate, the estimate in Eq. (6) prefers clusters of frames that use the same *categories* as context. While some of the results of these preferences will be the same, the latter approach lets the model make second-order inferences about categories. There may be no context words in common between the current frame and a potential cluster, but if the context words in the cluster have been found to be distributionally similar to those in the frame, it may be a good cluster for that frame.

We equally weight the word-based and the category-based estimates for $P(w_i|k)$ to get the likelihood of a context word; that is:

$$P(w_i|k) \approx \frac{1}{2}P_{word}(w_i|k) + \frac{1}{2}P_{cat}(w_i|k) \qquad (7)$$

This way, the model sees an input utterance simultaneously as a sequence of words and as a sequence of categories. It is the P_{cat} component, by using developing category knowledge, that yields the bootstrapping abilities of our model.

2.4 Generalization

Our model relies heavily on the similarity of word contexts in order to find category structure. In natural language, these context features are highly variable, so it is difficult to draw consistent structure from the input in the early stages of an incremental model. When little information is available, there is a risk of incorrectly generalizing, leading to clustering errors which may be difficult to overcome. Children face a similar problem in early learning, but there is evidence that they may manage the problem by using conservative strategies (see, e.g., Tomasello, 2000). Children may form specific hypotheses about each word type, only later generalizing their knowledge to similar words. Drawing on this observation, we form early small clusters specific to the head word type, then later aid generalization by merging these smaller clusters. By doing this, we ensure that the model only groups words of different types when there is sufficient evidence for their contextual similarity.

Thus, when a cluster has been newly created, we require that all frames put into the cluster share the same head word type.[2] When clusters are small, this prevents the model from making potentially incorrect generalizations to different words. Periodically, we evaluate a set of reasonably-sized clusters, and merge pairs of clusters that have highly similar contexts (see below for details). If the model decides to merge two clusters with different head word types—e.g., one cluster with all instances of *dog*, and another with *cat*—it has in effect made a decision to generalize. Intuitively, the model has learned that the contexts in the newly merged cluster apply to more than one word type. We now say that *any* word type could be a member of this cluster, if its context is sufficiently similar to that of the cluster. Thus, when categorizing a new word token (represented as a frame F), our model can choose from among the clusters with a matching head word, and any of these 'generalized' clusters that contain mixed head words.

Periodically, we look through a subset of the clusters to find similar pairs to merge. In order to limit the number of potential merges to consider, we only examine pairs of clusters in which at least one cluster has changed since the last check. Thus, after processing every 100 frames of input, we consider the clusters used to hold those recent 100 frames as candidates to be merged with another cluster. We only consider clusters of reasonable size (here, at least 10 frames) as candidates for merging. For each candidate pair of clusters, k_1 and k_2, we first evaluate a heuristic merge score that determines if the pair is appropriate to be merged, according to some local criteria, i.e., the size and the contents of the candidate clusters. For each suggested merge (a pair whose merge score exceeds a pre-determined threshold), we then look at the set of all clusters, the *global* evidence, to decide whether to accept the merge.

The merge score combines two factors: the entrenchment of the two clusters, and the similarity of

[2] However, a word type may exist in several clusters (e.g., for distinct noun and verb usages), thus handling lexical ambiguity.

their context features. The entrenchment measure identifies clusters that contain enough frames to show a significant trend. We take a sigmoid function over the number of frames in the clusters, giving a soft threshold approaching 0 for small clusters and 1 for large clusters. The similarity measure identifies pairs of clusters with similar distributions of word and category contexts. Given two clusters, we measure the symmetric Kullback-Leibler divergence for each corresponding pair of context feature probabilities (including the category contexts $P_i(k'|k)$, 8 pairs in total), then place the sum of those measures on another sigmoid function. The merge score is the sum of the entrenchment and similarity measures.

Since it is only a local measure, the merge score is not sufficient on its own for determining if a merge is appropriate. For each suggested merge, we thus examine the likelihood of a sample of input frames (here, the last 100 frames) under two states: the set of clusters before the merge, and the set of clusters if the merge is accepted. We only accept a merge if it results in an increase in the likelihood of the sample data. The likelihood of a sample set of frames, $\mathcal{S}$, over a set of clusters, $\mathcal{K}$, is calculated as in:

$$P(\mathcal{S}) = \prod_{F \in \mathcal{S}} \sum_{k \in \mathcal{K}} P(F|k)P(k) \quad (8)$$

3 Evaluation Methodology

To test our proposed model, we train it on a sample of language representative of what children would hear, and evaluate its categorization abilities. We have multiple goals in this evaluation. First, we determine the model's ability to discover adult-level syntactic categories from the input. Since this is intended to be a cognitively plausible learning model, we also compare the model's qualitative learning behaviours with those of children. In the first experiment (Section 4), we compare the model's categorization with a gold standard of adult-level syntactic categories and examine the effect of the bootstrapping component. The second experiment (Section 5) examines the model's development of three specific parts of speech. Developmental evidence suggests that children acquire different syntactic categories at different ages, so we compare the model's learning rates of nouns, verbs, and adjectives. Lastly, we examine our model's ability to handle lexically ambiguous words (Section 6). English word forms commonly belong to more than one syntactic category, so we show how our model uses context to disambiguate a word's category.

In all experiments, we train and test the model using the Manchester corpus (Theakston et al., 2001) from the CHILDES database (MacWhinney, 2000). The corpus contains transcripts of mothers' conversations with 12 British children between the ages of 1;8 (years;months) and 3;0. There are 34 one-hour sessions per child over the course of a year. The age range of the children roughly corresponds with the ages at which children show the first evidence of syntactic categories.

We extract the mothers' speech from each of the transcripts, then concatenate the input of all 12 children (all of Anne's sessions, followed by all of Aran's sessions, and so on). We remove all punctuation. We spell out contractions, so that each token in the input corresponds to only one part-of-speech (PoS) label (noun, verb, etc.). We also remove single-word utterances and utterances with a single repeated word type, since they contain no distributional information. We randomly split the data into development and evaluation sets, each containing approximately 683,000 tokens. We use the development set to fine-tune the model parameters and develop the experiments, then use the evaluation set as a final test of the model. We further split the development set into about 672,000 tokens (about 8,000 types) for training and 11,000 tokens (1,300 types) for validation. We split the evaluation set comparably, into training and test subsets. All reported results are for the evaluation set. A conservative estimate suggests that children are exposed to at least 1.5 million words of child-directed speech annually (Redington et al., 1998), so this corpus represents only a small portion of a child's available input.

4 Experiment 1: Adult Categories

4.1 Methods

We use three separate versions of the categorization model, in which we change the components used to estimate the context word probability, $P(w_i|k)$ (as used in Eq. (5), Section 2.2). In the *word-based* model, we estimate the context probabilities using only the words in the context window, by directly using the maximum-likelihood P_{word} estimate. The *bootstrap* model uses only the existing clusters to estimate the probability, directly using the P_{cat} estimate from Eq. (6). The *combination* model uses an equally-weighted combination of the two probabilities, as presented in Eq. (7).

We run the model on the training set, categorizing each of the resulting frames in order. After every 10,000 words of input, we evaluate the model's categorization performance on the test set. We categorize each of the frames of the test set as usual, treating the text as regular input. So that the test set remains unseen, the model does not record these categorizations.

4.2 Evaluation

The PoS tags in the Manchester corpus are too fine-grained for our evaluation, so for our gold standard

we map them to the following 11 tags: noun, verb, auxiliary, adjective, adverb, determiner, conjunction, negation, preposition, infinitive *to*, and 'other.' When we evaluate the model's categorization performance, we have two different sets of clusters of the words in the test set: one set resulting from the gold standard, and another as a result of the model's categorization. We compare these two clusterings, using the adjusted Rand index (Hubert and Arabie, 1985), which measures the overall agreement between two clusterings of a set of data points. The measure is 'corrected for chance,' so that a random grouping has an expected score of zero. This measure tends to be very conservative, giving values much lower than an intuitive percentage score. However, it offers a useful relative comparison of overall cluster similarity.

4.3 Results

Figure 1 gives the adjusted Rand scores of the three model variants, *word-based*, *bootstrap*, and *combination*. Higher values indicate a better fit with the gold-standard categorization scheme. The adjusted Rand score is corrected for chance, thus providing a built-in baseline measure. Since the expected score for a random clustering is zero, all three model variants operate at above-baseline performance.

As seen in Figure 1, the word-based model gains an early advantage in the comparison, but its performance approaches a plateau at around 200,000 words of input. This suggests that while simple word distributions provide a reliable source of information early in the model's development, the information is not sufficient to sustain long-term learning. The bootstrap model learns much more slowly, which is unsurprising, given that it depends on having some reasonable category knowledge in order to develop its clusters—leading to a chicken-and-egg problem. However, once started, its performance improves well beyond the word-based model's plateau. These results suggest that on its own, each component of the model may be effectively throwing away useful information. By combining the two models, the combination model appears to gain complementary benefits from each component, outperforming both. The word-based component helps to create a base of reliable clusters, which the bootstrap component uses to continue development.

After all of the training text, the combination model uses 411 clusters to categorize the test tokens (compared to over 2,000 at the first test point). While this seems excessive, we note that 92.5% of the test tokens are placed in the 25 most populated clusters.[3]

[3]See www.cs.toronto.edu/~chris/syncat for examples.

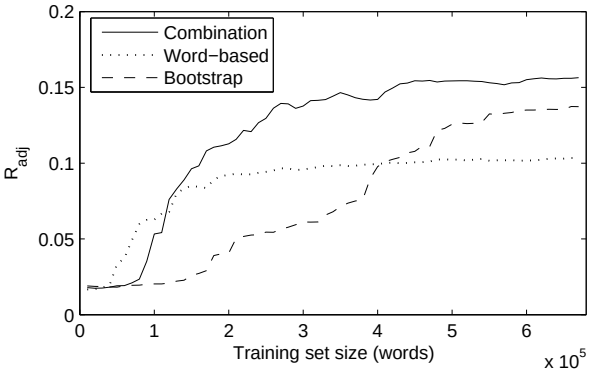

Figure 1: Adjusted Rand Index of each of three models' clusterings of the test set, as compared with the PoS tags of the test data.

5 Experiment 2: Learning Trends

A common trend observed in children is that different syntactic categories are learned at different rates. Children appear to have learned the category of nouns by 23 months of age, verbs shortly thereafter, and adjectives relatively late (Kemp et al., 2005). Our goal in this experiment is to look for these specific trends in the behaviour of our model. We thus simulate an experiment where a child uses a novel word's linguistic context to infer its syntactic category (e.g., Tomasello et al., 1997). For our experiment, we randomly generate input frames with novel head words using contexts associated with nouns, verbs, and adjectives, then examine the model's categorization in each case. We expect that our model should approximate the developmental trends of children, who tend to learn the category of 'noun' before 'verb,' and both of these before 'adjective.'

5.1 Methods

We generate new input frames using the most common syntactic patterns in the training data. For each of the noun, verb, and adjective categories (from the gold standard), we collect the five most frequent PoS sequences in which these are used, bounded by the usual four-word context window. For example, the Adjective set includes the sequence 'V Det **Adj** N *null*', where the sentence ends after the N. For each of the three categories, we generate each of 500 input frames by sampling one of the five PoS sequences, weighted by frequency, then sampling words of the right PoS from the lexicon, also weighted by frequency. We replace the head word with a novel word, forcing the model to use only the context for clustering. Since the context words are chosen at random, most of the word sequences generated will be novel. This makes the task more difficult, rather than simply sampling utterances from the corpus, where rep-

etitions are common. While a few of the sequences may exist in the training data, we expect the model to mostly use the underlying category information to cluster the frames.

We intend to show that the model uses context to find the right category for a novel word. To evaluate the model's behaviour, we let it categorize each of the randomly generated frames. We score each frame as follows: if the frame gets put into a new cluster, it earns score zero. Otherwise, its score is the proportion of frames in the chosen cluster matching the correct part of speech (we use a PoS-tagged version of the training corpus; for example, a noun frame put into a cluster with 60% nouns would get 0.6). We report the mean score for each of the noun, verb, and adjective sets. Intuitively, the matching score indicates how well the model recognizes that the given contexts are similar to input it has seen before. If the model clusters the novel word frame with others of the right type, then it has formed a category for the contextual information in that frame.

We use the full combination model (Eq. (7)) to evaluate the learning rates of individual parts of speech. We run the model on the training subset of the evaluation corpus. After every 10,000 words of input, we use the model to categorize the 1,500 context frames with novel words (500 frames each for noun, verb, and adjective). As in experiment 1, the model does not record these categorizations.

5.2 Results

Figure 2 shows the mean matching scores for each of the tested parts of speech. Recall that since the frames each use a novel head word, a higher matching score indicates that the model has learned to correctly recognize the contexts in the frames. This does not necessarily mean that the model has learned single, complete categories of 'noun,' 'verb,' and 'adjective,' but it does show that when the head word gives no information, the model can generalize based on the contextual patterns alone. The model learns to categorize novel nouns better than verbs until late in training, which matches the trends seen in children. Adjectives progress slowly, and show nearly no learning ability by the end of the trial. Again, this appears to reflect natural behaviour in children, although the effect we see here may simply be a result of the overall frequency of the PoS types. Over the entire corpus (development and evaluation), 35.4% of the word tokens are nouns and 24.3% are verbs, but only 2.9% are tagged as adjectives. The model, and similarly a child, may need much more data to learn adjectives than is available at this stage.

The scores in Figure 2 tend to fluctuate, particularly for the noun contexts. This fluctuation corresponds to periods of overgeneralization, followed by recovery (also observed in children; see, e.g., Tomasello, 2000). When the model merges two clusters, the contents of the resulting cluster can initially be quite heterogeneous. Furthermore, the new cluster is much larger, so it becomes a magnet for new categorizations. This results in overgeneralization errors, giving the periodic drops seen in Figure 2. While our formulation in Section 2.4 aims to prevent such errors, they are likely to occur on occasion. Eventually, the model recovers from these errors, and it is worth noting that the fluctuations diminish over time. As the model gradually improves with more input, the dominant clusters become heavily entrenched, and inconsistent merges are less likely to occur.

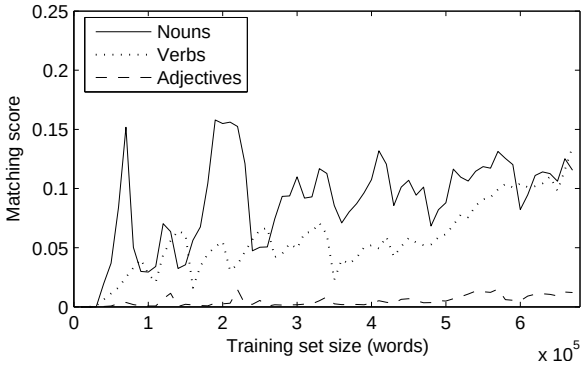

Figure 2: Comparative learning trends of noun, verb, and adjective patterns.

6 Experiment 3: Disambiguation

The category structure of our model allows a single word type to be a member of multiple categories. For example, *kiss* could belong to a category of predominantly noun usages (*Can I have a kiss?*) and also to a category of verb usages (*Kiss me!*). As a result, the model easily represents lexical ambiguity. In this experiment, inspired by disambiguation work in psycholinguistics (see, e.g., MacDonald, 1993), we examine the model's ability to correctly disambiguate category memberships.

6.1 Methods

Given a word that the model has previously seen as various different parts of speech, we examine how well the model can use that ambiguous word's context to determine its category in the current usage. For example, by presenting the word *kiss* in separate noun and verb contexts, we expect that the model should categorize *kiss* as a noun, then as a verb, respectively. We also wish to examine the effect of the target word's lexical bias, that is, the predominance of a word type to be used as one category over another. As with adults, if *kiss* is mainly used as a noun, we expect the model to more accurately categorize the

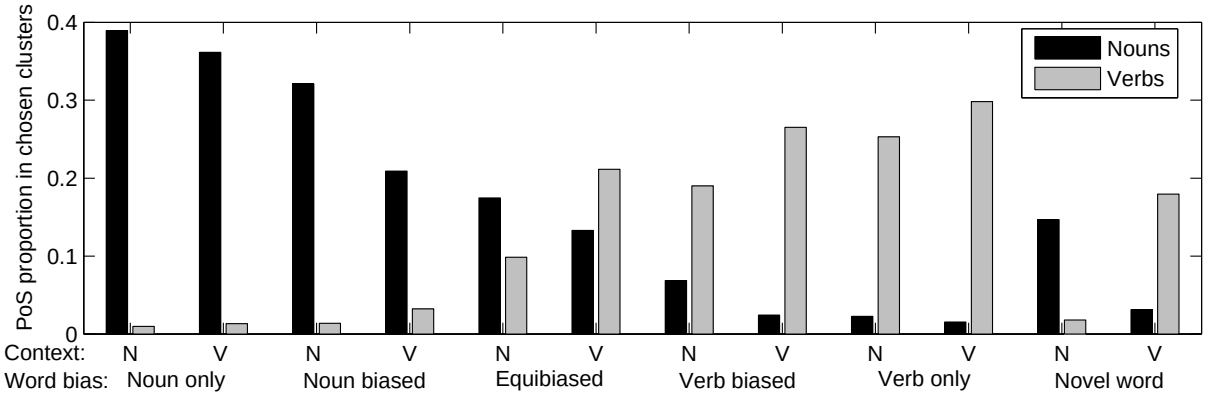

Figure 3: Syntactic category disambiguation. Shown are the proportions of nouns and verbs in the chosen clusters for ambiguous words used in either noun (N) or verb (V) contexts.

word in a noun context than in a verb context.

We focus on noun/verb ambiguities. We artificially generate input frames for noun and verb contexts as in experiment 2, with the following exceptions. To make the most use of the context information, we allow no *null* words in the input frames. We also ensure that the contexts are distinctive enough to guide disambiguation. For each PoS sequence surrounding a noun (e.g., 'V Det **head** Prep Det'), we ensure that over 80% of the instances of that pattern in the corpus are for nouns, and likewise for verbs.

We test the model's disambiguation in six conditions, with varying degrees of lexical bias. Unambiguous ('noun/verb only') conditions test words seen in the corpus only as nouns or verbs (10 words each). 'Biased' conditions test words with a clear bias (15 with average 93% noun bias; 15 with average 84% verb bias). An 'equibiased' condition uses 4 words of approximately equal bias, and a novel word condition provides an unbiased case.

For the six sets of test words, we measure the effect of placing each of these words in both noun and verb contexts. That is, each word in each condition was used as the head word in each of the 500 noun and 500 verb disambiguating frames. For example, we create 500 frames where *book* is used as a noun, and 500 frames where it is used as a verb. We then use the fully-trained 'combination' model (Eq. (7)) to categorize each frame. Unlike in the previous experiment, we do not let the model create new clusters. For each frame, we choose the best-fitting existing cluster, then examine that cluster's contents. As in experiment 2, we measure the proportions of each PoS of the frames in this cluster. We then average these measures over all tested frames in each condition.

6.2 Results

Figure 3 presents the measured PoS proportions for each of the six conditions. For both the equibias and novel word conditions, we see that the clusters chosen for the noun context frames (labeled N) contain more nouns than verbs, and the clusters chosen for the verb context frames (V) contain more verbs than nouns. This suggests that although the model's past experience with the head word is not sufficiently informative, the model can use the word's context to disambiguate its category. In the 'unambiguous' and the 'biased' conditions, the head words' lexical biases are too strong for the model to overcome.

However, the results show a realistic effect of the lexical bias. Note the contrasts from the 'noun only' condition, to the 'noun biased' condition, to 'equibiased' (and likewise for the verb biases). As the lexical bias weakens, the counter-bias contexts (e.g., a noun bias with a verb context) show a stronger effect on the chosen clusters. This is a realistic effect of disambiguation seen in adults (MacDonald, 1993). Strongly biased words are more difficult to categorize in conflict with their bias than weakly biased words.

7 Related Work

Several existing computational models use distributional cues to find syntactic categories. Schütze (1993) employs co-occurrence statistics for common words, while Redington et al. (1998) build word distributional profiles using corpus bigram counts. Clark (2000) also builds distributional profiles, introducing an iterative clustering method to better handle ambiguity and rare words. Mintz (2003) shows that even very simple three-word templates can effectively define syntactic categories. Each of these models demonstrates that by using the kinds of simple information to which children are known to be sensitive, syntactic categories are learnable. However, the specific learning mechanisms they use, such as the hierarchical clustering methods of Redington et al. (1998), are not intended to be cognitively plausible.

In contrast, Cartwright and Brent (1997) propose

an incremental model of syntactic category acquisition that uses a series of linguistic preferences to find common patterns across sentence-length templates. Their model presents an important incremental algorithm which is very effective for discovering categories in artificial languages. However, the model's reliance on templates limits its applicability to transcripts of actual spoken language data, which contain high variability and noise.

Recent models that apply Bayesian approaches to PoS tagging are not incremental and assume a fixed number of tags (Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008). In syntactic category acquisition, the true number of categories is unknown, and must be inferred from the input.

8 Conclusions and Future Directions

We have developed a computational model of syntactic category acquisition in children, and demonstrated its behaviour on a corpus of naturalistic child-directed data. The model is based on domain-general properties of feature similarity, in contrast to earlier, more linguistically-specific methods. The incremental nature of the algorithm contributes to a substantial improvement in psychological plausibility over previous models of syntactic category learning. Furthermore, due to its probabilistic framework, our model is robust to noise and variability in natural language.

Our model successfully uses a syntactic bootstrapping mechanism to build on the distributional properties of words. Using its existing partial knowledge of categories, the model applies a second level of analysis to learn patterns in the input. By making few assumptions about prior linguistic knowledge, the model develops realistic syntactic categories from the input data alone. The explicit bootstrapping component improves the model's ability to learn adult categories, and its learning trajectory resembles relevant behaviours seen in children. Using the contextual patterns of individual parts of speech, we show differential learning rates across nouns, verbs, and adjectives that mimic child development. We also show an effect of a lexical bias in category disambiguation.

The algorithm is currently only implemented as an incremental process. However, comparison with a batch version of the algorithm, such as by using a Gibbs sampler (Sanborn et al., 2006), would help us further understand the effect of incrementality on language fidelity.

While we have only examined the effects of learning categories from simple distributional information, the feature-based framework of our model could easily be extended to include other sources of information, such as morphological and phonological cues. Furthermore, it would also be possible to include semantic features, thereby allowing the model to draw on correlations between semantic and syntactic categories in learning.

Acknowledgments

We thank Afra Alishahi for valuable discussions, and the anonymous reviewers for their comments. We gratefully acknowledge the financial support of NSERC of Canada and the University of Toronto.

References

Alishahi, A. and S. Stevenson 2008. A computational model for early argument structure acquisition. *Cognitive Science*, 32(5).

Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Cartwright, T. A. and M. R. Brent 1997. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, 63:121–170.

Clark, A. 2000. Inducing syntactic categories by context distribution clustering. In *CoNLL2000*, pp. 91–94.

Goldwater, S. and T. L. Griffiths 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proc. of ACL2007*, pp. 744–751.

Hubert, L. and P. Arabie 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Kemp, N., E. Lieven, and M. Tomasello 2005. Young children's knowledge of the "determiner" and "adjective" categories. *J. Speech Lang. Hear. R.*, 48:592–609.

MacDonald, M. C. 1993. The interaction of lexical and syntactic ambiguity. *J. Mem. Lang.*, 32:692–715.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*, volume 2: The Database. Lawrence Erlbaum, Mahwah, NJ, 3 edition.

Mintz, T. H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90:91–117.

Olguin, R. and M. Tomasello 1993. Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8:245–272.

Onnis, L. and M. H. Christiansen 2005. New beginnings and happy endings: psychological plausibility in computational models of language acquisition. *CogSci2005*.

Redington, M., N. Chater, and S. Finch 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Sanborn, A. N., T. L. Griffiths, and D. J. Navarro 2006. A more rational model of categorization. *CogSci2006*.

Schütze, H. 1993. Part of speech induction from scratch. In *Proc. of ACL1993*, pp. 251–258.

Theakston, A. L., E. V. Lieven, J. M. Pine, and C. F. Rowland 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *J. Child Lang.*, 28:127–152.

Tomasello, M. 2000. Do young children have adult syntactic competence? *Cognition*, 74:209–253.

Tomasello, M., N. Akhtar, K. Dodson, and L. Rekau 1997. Differential productivity in young children's use of nouns and verbs. *J. Child Lang.*, 24:373–387.

Toutanova, K. and M. Johnson 2008. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *NIPS2008*.