Online Entropy-based Model of Lexical Category Acquisition

Grzegorz Chrupała
Saarland University
gchrupala@lsv.uni-saarland.de

Afra Alishahi
Saarland University
afra@coli.uni-saarland.de

Abstract

Children learn a robust representation of lexical categories at a young age. We propose an incremental model of this process which efficiently groups words into lexical categories based on their local context using an information-theoretic criterion. We train our model on a corpus of child-directed speech from CHILDES and show that the model learns a fine-grained set of intuitive word categories. Furthermore, we propose a novel evaluation approach by comparing the efficiency of our induced categories against other category sets (including traditional part of speech tags) in a variety of language tasks. We show the categories induced by our model typically outperform the other category sets.

1 The Acquisition of Lexical Categories

Psycholinguistic studies suggest that early on children acquire robust knowledge of the abstract lexical categories such as nouns, verbs and determiners (e.g., Gelman & Taylor, 1984; Kemp et al., 2005). Children's grouping of words into categories might be based on various cues, including phonological and morphological properties of a word, the distributional information about its surrounding context, and its semantic features. Among these, the distributional properties of the local context of a word have been thoroughly studied. It has been shown that child-directed speech provides informative co-occurrence cues, which can be reliably used to form lexical categories (Redington et al., 1998; Mintz, 2003).

The process of learning lexical categories by children is necessarily incremental. Human language acquisition is bounded by memory and processing limitations, and it is implausible that humans process large volumes of text at once and induce an optimum set of categories. Efficient online computational models are needed to investigate whether distributional information is equally useful in an online process of word categorization. However, the few incremental models of category acquisition which have been proposed so far are generally inefficient and over-sensitive to the properties of the input data (Cartwright & Brent, 1997; Parisien et al., 2008). Moreover, the unsupervised nature of these models makes their assessment a challenge, and the evaluation techniques proposed in the literature are limited.

The main contributions of our research are twofold. First, we propose an incremental entropy model for efficiently clustering words into categories given their local context. We train our model on a corpus of child-directed speech from CHILDES (MacWhinney, 2000) and show that the model learns a fine-grained set of intuitive word categories. Second, we propose a novel evaluation approach by comparing the efficiency of our induced categories against other category sets, including the traditional part of speech tags, in a variety of language tasks. We evaluate our model on word prediction (where a missing word is guessed based on its sentential context), semantic inference (where the semantic properties of a novel word are predicted based on the context), and grammaticality judgment (where the syntactic well-formedness of a sentence is assessed based on the category labels assigned to its words). The results show that the categories induced by our model can be successfully used in a variety of tasks and typically perform better than other category sets.

1.1 Unsupervised Models of Category Induction

Several computational models have used distributional information for categorizing words (e.g. Brown et al., 1992; Redington et al., 1998; Clark, 2000; Mintz, 2002). The majority of these mod-

els partition the vocabulary into a set of optimum clusters (e.g., Brown et al., 1992; Clark, 2000). The generated clusters are intuitive, and can be used in different tasks such as word prediction and parsing. Moreover, these models confirm the learnability of abstract word categories, and show that distributional cues are a useful source of information for this purpose. However, (i) they categorize word types rather than word tokens, and as such provide no account of words belonging to more than one category, and (ii) the batch algorithms used by these systems make them implausible for modeling human category induction. Unsupervised models of PoS tagging such as Goldwater & Griffiths (2007) do assign labels to word-tokens, but they still typically use batch processing, and what is even more problematic, they hardwire important aspects of the model, such as the final number of categories.

Only few previously proposed models process data incrementally, categorize word-tokens and do not pre-specify a fixed category set. The model of Cartwright & Brent (1997) uses an algorithm which incrementally merges word clusters so that a Minimum Description Length criterion for a template grammar is optimized. The model treats whole sentences as contextual units, which sacrifices a degree of incrementality, as well as making it less robust to noise in the input.

Parisien et al. (2008) propose a Bayesian clustering model which copes with ambiguity and exhibits the developmental trends observed in children (e.g. the order of acquisition of different categories). However, their model is overly sensitive to context variability, which results in the creation of sparse categories. To remedy this issue they introduce a "bootstrapping" component where the categories assigned to context words are use to determine the category of the current target word. They also perform periodical cluster reorganization. These mechanisms improve the overall performance of the model when trained on large amounts of training data, but they complicate the model with ad-hoc extensions and add to the (already considerable) computational load.

What is lacking is an incremental model of lexical category which can efficiently process naturalistic input data and gradually build robust categories with little training data.

1.2 Evaluation of the Induced Categories

There is no standard and straightforward method for evaluating the unsupervised models of category learning (see Clark, 2003, for discussion). Many unsupervised models of lexical category acquisition treat the traditional part of speech (PoS) tags as the gold standard, and measure the accuracy and completeness of their induced categories based on how closely they resemble the PoS categories (e.g. Redington et al., 1998; Mintz, 2003; Parisien et al., 2008). However, it is not at all clear whether humans form the same types of categories. In fact, many language tasks might benefit from finer-grained categories than the traditional PoS tags used for corpus annotation.

Frank et al. (2009) propose a different, automatically generated set of gold standard categories for evaluating an unsupervised categorization model. The gold-standard categories are formed according to "substitutability": if one word can be replaced by another and the resulting sentence is still grammatical, then there is a good chance that the two words belong to the same category. They extract 3-word frames from the training data, and form the gold standard categories based on the words that appear in the same frame. They emphasize that in order to provide some degree of generalization, different data sets must be used for forming the gold-standard categories and performing the evaluation. However, the resulting categories are bound to be incomplete, and using them as gold standard inevitably favors categorization models which use a similar frame-based principle.

All in all, using any set of gold standard categories for evaluating an unsupervised categorization model has the disadvantage of favoring one set of principles and intuitions over another; that is, assuming that there is a *correct* set of categories which the model should converge to. Alternatively, automatically induced categories can be evaluated based on how *useful* they are in performing different tasks. This approach is taken by Clark (2000), where the perplexity of a finite-state model is used to compare different category sets.

We build on this idea and propose a more general usage-based approach to evaluating the automatically induced categories from a data set, emphasizing that the ultimate goal of a category induction model is to form categories that can be efficiently used in a variety of language tasks. We argue that for such tasks, a finer-grained set of cat-

egories might be more appropriate than the coarse-grained PoS categories. Therefore, we propose a number of tasks for which we compare the performance based on various category sets, including those induced by our model.

2 An Incremental Entropy-based Model of Category Induction

A model of human category acquisition should possess two key features:

- It should process input as it arrives, and incrementally update the current set of clusters.
- The set of clusters should not be fixed in advance, but rather determined by the characteristics of the input data.

We propose a simple algorithm which fulfills those two conditions.

Our goal is to categorize word usages based on the similarity of their form (the content) and their surrounding words (the context). While grouping word usages into categories, we attempt to trade off two conflicting criteria. First, the categories should be informative about the properties of their members. Second, the number and distribution of the categories should be parsimonious. An appropriate tool for formalizing both informativeness and parsimony is information-theoretic entropy.

The parsimony criterion can be formalized as the entropy of the random variable (Y) representing the cluster assignments:

$$H(Y) = -\sum_{i=1}^{N} P(Y = y_i) \log_2(P(Y = y_i)) \quad (1)$$

where N is the number of clusters and $P(Y = y_i)$ stands for the relative size of the i^{th} cluster.

The informativeness criterion can be formalized as the conditional entropy of training examples (X) given the cluster assignments:

$$H(X|Y) = \sum_{i=1}^{N} P(Y = y_i) H(X|Y = y_i) \quad (2)$$

and $H(X|Y = y_i)$ is calculated as

$$H(X|Y = y_i) = -\sum_{j=1}^{T} [P(X = x_j|Y = y_i) \times \log_2(P(X = x_j|Y = y_i)] \quad (3)$$

where T is the number of word usages in the training set.

The two criteria presented by Equations 1 and 2 can be combined together as the joint entropy of the two random variables X and Y:

$$H(X, Y) = H(X|Y) + H(Y) \quad (4)$$

For a random variable X corresponding to a single feature, minimizing the joint entropy $H(X, Y)$ will trade off our two desired criteria.

The joint entropy will be minimal if each distinct value of variable X is assigned the same category (i.e. same value of Y). There are many assignments which satisfy this condition. They range from putting all values of X in a single category, to having a unique category for each unique value of X. We favor the latter solution algorithmically by creating a new category in case of ties.

Finally, since our training examples contain a bundle of categorical features, we minimize the joint entropy simultaneously for all the features. We consider our training examples to be vectors of random variables $(X_j)_{j=1}^{M}$, where each random variable corresponds to one feature. For an incoming example we will choose the cluster assignment which leads to the least increase in the joint entropy $H(X_j, Y)$, summed over all the features j:

$$\begin{aligned} \sum_{j=1}^{M} H(X_j, Y) &= \sum_{j=1}^{M} \big[H(X_j|Y) + H(Y)\big] \quad (5) \\ &= \sum_{j=1}^{M} \big[H(X_j|Y)\big] + M \times H(Y) \end{aligned}$$

In the next section, we present an incremental algorithm which uses this criterion for inducing categories from a sequence of input data.

The Incremental Algorithm. For each word usage that the model processes at time t, we need to find the best category among the ones that have been formed so far, as well as a potential new category. The decision is made based on the change in the function $\sum_{j=1}^{M} H(X_j, Y)$ (Equation 5) from point $t-1$ to point t, as a result of assigning the current input x^t to a category y:

$$\Delta H_y^t = \sum_{j=1}^{M} \big[H_y^t(X_j, Y) - H^{t-1}(X_j, Y)\big] \quad (6)$$

where $H_y^t(X, Y)$ is the joint entropy of the assignment Y for the input $X = \{x^1, \ldots, x^t\}$, *after* the last input item x^t is assigned to the category y. The winning category $\hat{y}$ is the one that leads to the smallest increase. Ties are broken by preferring a new category.

$$\hat{y} = \begin{cases} \text{argmin}_{y \in \{y\}_{i=1}^{N}} \Delta H_y^t & \text{if } \exists y_n [\Delta H_{y_n}^t < \Delta H_{y_{N+1}}^t] \\ y_{N+1} & \text{otherwise} \end{cases} \quad (7)$$

where N is the number of categories created up to point t, and y_{N+1} represents a new category.

Efficiency. We maintain the relative size $P^t(y)$ and the entropy $H(X_j|Y = y)$ for each category y over time. When performing an assignment of x^t to a category y_i, we only need to update the conditional entropies $H(X_j|Y = y_i)$ for all features X_j for this particular category, since other categories have not changed. For a feature X_j at point t, the change in the conditional entropy for the selected category y_i is given by:

$$\begin{aligned}\Delta H^t_{y_i}(X_j|Y) &= H^t_{y_i}(X_j|Y) - H^{t-1}(X_j|Y) \\ &= \sum_{y_k \neq y_i} \left[P(Y = y_k)H^{t-1}(X_j|Y = y_i)\right] \\ &\quad - P^{t-1}(Y = y_i)H^{t-1}(X|Y = y_i) \\ &\quad - P^t(Y = y_i)H^t(X_j|Y = y_i)\end{aligned}$$

where only the last term depends on the current time index t. Therefore, the entropy $H(X_j|Y)$ at each step can be efficiently updated by calculating this term for the modified category at that step.

A number of previous studies have considered entropy-based criteria for clustering (e.g. Barbará et al., 2002; Li et al., 2004). The main contribution of our proposed model is the emphasis on rarely explored combination of the two characteristics we consider crucial for modeling human category acquisition, incrementality and an open set of clusters.

3 Experimental Setup

We evaluate the categories formed by our model through three different tasks. The first task is word prediction, where a target word is predicted based on the sentential context it appears in. The second task is to infer the semantic properties of a novel word based on its context. The third task is to assess the grammaticality of a sentence tagged with category labels. We run our model on a corpus of child-directed speech, and use the categories that it induces from that corpus in the above-mentioned tasks. For each task, we compare the performance using our induced categories against the performance using other category sets. In the following sections, we describe the properties of the data sets used for training and testing the model, and the formation of other category sets against which we compare our model.

Data Set	Sessions	#Sentences	#Words
Training	26–28	22,491	125,339
Development	29–30	15,193	85,361
Test	32–33	14,940	84,130

Table 1: Experimental data

3.1 Input Data

We use the Manchester corpus (Theakston et al., 2001) from CHILDES database (MacWhinney, 2000) as experimental data. The Manchester corpus consists of conversations with 12 children between the ages of eighteen months to three years old. The corpus is manually tagged using 60 PoS labels. We use the mother's speech from transcripts of 6 children, remove punctuation, and concatenate the corresponding sessions.

We used data from three sessions as the training set, two sessions as the development set, and two sessions as the test set. We discarded all one-word sentences from the data sets, as they do not provide any context for our evaluation tasks. Table 1 summarizes the properties of each data set.

3.2 Category Sets

We define each word usage in the training or test data set as a vector of three categorical features: the content feature (i.e., the focus word in a usage), and two context features (i.e. the preceding and following bigrams). We ran our clustering algorithm on the training set, which resulted in a set of 944 categories (of which 442 have only one member). Table 3 shows two sample categories from the training set, and Figure 1 shows the size distribution of the categories.

For each evaluation task, we use the following category sets to label the test set:

ΔH. The categories induced by our entropy-based model from the training set, as described above.

PoS. The part-of-speech tags the Manchester corpus is annotated with.

Words. The set of all the word types in the data set (i.e. assuming that all the usages of the same word form are grouped together).

Parisien. The induced categories by the model of Parisien et al. (2008) from the training set.

	Gold PoS	Words	Parisien	ΔH
VI	(0.000)	5.294	5.983	**4.806**
ARI	(1.000)	0.139	0.099	**0.168**

Table 2: Comparison against gold PoS tags using Variation of Information (VI) and Adjusted Rand Index (ARI).

Sample Cluster 1		Sample Cluster 2	
going	(928)	than	(45)
doing	(190)	more	(20)
back	(150)	silly	(10)
coming	(80)	bigger	(9)
looking	(76)	frightened	(5)
making	(64)	dark	(4)
playing	(55)	harder	(4)
taking	(45)	funny	(3)
...		...	

Table 3: Sample categories induced from the training data. The frequency of each word in the category is shown in parentheses.

For the first two tasks (word prediction and semantic inference), we do not use the content feature in labeling the test set, since the assumption underlying both tasks is that we do not have access to the form of the target word. Therefore, we do not measure the performance of these tasks on the **Words** category set. However, we do use the content feature in labeling the test examples in grammaticality judgment.

For completeness, in Table 2 we report the results of evaluation against Gold PoS tags using two metrics, Variation of Information (Meila, 2003) and Adjusted Rand Index (Hubert & Arabie, 1985).

4 Word Prediction

Humans can predict a word based on the context it is used in with remarkable accuracy (e.g. Lesher et al., 2002). Different versions of this task such as Cloze Test (Taylor, 1953) are used for the assessment of native and second language learning.

We simulate this task, where a missing word is predicted based on its context. We use each of the category sets introduced in Section 3.2 to label a word usage in the test set, without using the word form itself as a feature. That is, we assume that the target word is unknown, and find the best category for it based only on its surrounding context.

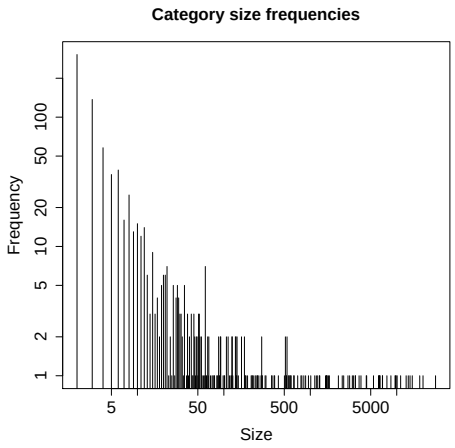

Figure 1: The distribution of the induced categories based on their size

We then output a ranked list of the content feature values of the selected category as the prediction of the model for the target word. To evaluate this prediction, we use the reciprocal rank of the target word in the predicted list.

The third row of Table 4 shows the Mean Reciprocal Rank (MRR) over all the word usages in the test data across different category sets. The results show that the category labels predicted by our model (ΔH) perform much better than those of Parisien, but still not as good as the gold-standard part of speech categories. The fact that PoS tags are better here does not necessarily mean that the PoS category set is better for word prediction as such, since they are manually assigned and thus noise-free, unlike the automatic category labels predicted by the two models. In the second set of experiments described below we try to factor in the uncertainty about category assignment inherent in automatic labeling.

Using only the best category output by the model to produce word predictions is simple and neutral; however, it discards part of the information learned by the model. We can predict words more accurately by combining information from the whole ranked list of category labels.

We use the ΔH model to rank the values of the content feature in the following fashion: for the current test usage, we rank each cluster assignment y by the change in the $\Delta H^t_{y_i}$ function that it causes. For each of the assignments, we compute the relative frequencies $P(w|y_i)$ of each possible focus word. The final rank of the word w in context h is determined by the sum of the cluster-

	Gold PoS	Words	Parisien	ΔH
Word Prediction (MRR)	**0.354**	-	0.212	0.309
Semantic Inference (MAP)	0.351	-	0.213	**0.366**
Grammaticality Judgment (Accuracy)	**0.728**	0.685	0.683	0.715

Table 4: The performance in each of the three tasks using different category sets.

dependent relative frequencies weighted by the normalized reciprocal ranks of the clusters:

$$P(w|h) = \sum_{i=1}^{N} P(w|y_i) \frac{\mathrm{R}(y_i|h)^{-1}}{\sum_{i=1}^{N} \mathrm{R}(y_i|h)^{-1}} \quad (8)$$

where $\mathrm{R}(y_i|h)^{-1}$ is the reciprocal rank of cluster y_i for context h according to the model.

We compare the performance of the ΔH model with this word-prediction method to that of an n-gram language model, which is an established technique for assigning probabilities to words based on their context. For the language model we use several n-gram orders ($n = 1 \ldots 5$), and smooth the n-gram counts using absolute discounting (Zhai & Lafferty, 2004). The probability of the word w given the context h is given by the following model of order n:

$$P_n(w|h) = \max\left(0, \frac{\mathrm{c}(h,w) - d}{\mathrm{c}(h)}\right) + \alpha(h) P_{n-1}(w|h) \quad (9)$$

where d is the discount parameter, $\mathrm{c}(\cdot)$ is the frequency count function, P_{n-1} is the lower-order back-off distribution, and α is the normalization factor:

$$\alpha(h) = \begin{cases} 1 & \text{if } \mathrm{r}(h) = 0 \\ d\mathrm{r}(h)\frac{1}{\mathrm{c}(h)} & \text{otherwise} \end{cases} \quad (10)$$

and $\mathrm{r}(h)$ is the number of distinct words that follow context h in the training corpus.

In addition to the ΔH model and the n-gram language models, we also report how well words can be predicted from their manually assigned PoS tags from CHILDES: for each token we predict the most likely word given the token's true PoS tag based on frequencies in the training data.

Table 4 summarizes the evaluation results. The ΔH model can predict missing words better than any of the n-gram language models, and even slightly better than the true POS tags. Given the simplicity of our clustering model, this is a very encouraging result. Simple n-gram language models are known for providing quite a strong baseline for word prediction; for example, Brown et al. (1992)'s class-based language model failed to improve test-set perplexity over a word-based trigram model.

Model	MRR
LM $n = 1$	0.1253
LM $n = 2$	0.2884
LM $n = 3$	0.3278
LM $n = 4$	0.3305
LM $n = 5$	0.3297
ΔH	**0.3591**
Gold POS	0.3540

Table 5: Mean reciprocal rank on the word prediction task on the test set

5 Semantic Inference

Several experimental studies have shown that children and adults can infer (some aspects of) the semantic properties of a novel word based on the context it appears in (e.g. Landau & Gleitman, 1985; Gleitman, 1990; Naigles & Hoff-Ginsberg, 1995). For example, in an experimental study by Fisher et al. (2006), two-year-olds watched as a hand placed a duck on a box, and pointed to it as a new word was uttered. Half of the children heard the word presented as a noun (*This is a corp!*), while half heard it as a preposition (*This is acorp my box!*). After training, children heard a test sentence (*What else is acorp (my box)?*) while watching two test events: one showed another duck beside the box, and the other showed a different object on the box. Looking-preferences revealed effects of sentence context: subjects in the preposition condition interpreted the novel word as a location, whereas those in the noun condition interpreted it as an object.

To study a similar effect in our model, we associate each word with a set of semantic features. For nouns, we extract the semantic features from WordNet 3.0 (Fellbaum, 1998) as follows: We take all the hypernyms of the first sense of the word, and the first word in the synset of each hypernym to the set of the semantic features of

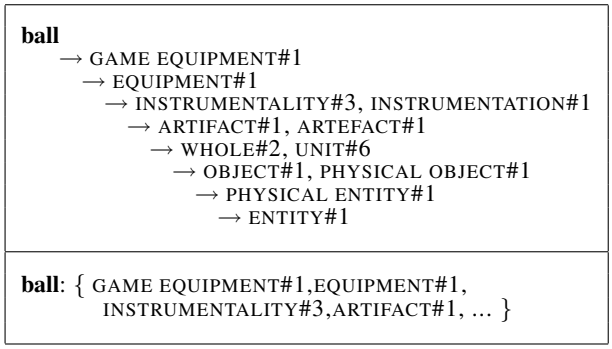

Figure 2: Semantic features of *ball*, as extracted from WordNet.

the target word (see Figure 2 for an example). For verbs, we additionally extract features from a verb-specific resource, VerbNet 2.3 (Schuler, 2005). Due to lack of proper resources for other lexical categories, we limit our evaluation to nouns and verbs.

The semantic features of words are not used in the formation of lexical categories. However, at each point of time in learning, we can associate a *semantic profile* to a category as the aggregated set of the semantic features of its members: each feature in the set is assigned a count that indicates the number of the category members which have that semantic property. This is done for each of the category sets described in Section 3.2.

As in the word-prediction task, we use different category sets to label each word usage in a test set based only on the context features of the word. When the model encounters a novel word, it can use the semantic profile of the word's labeled category as a prediction of the semantic properties of that word. We can evaluate the quality of this prediction by comparing the *true* meaning representation of the target word (i.e., its set of semantic features according to the lexicon) against the semantic profile of the selected category. We use the Mean Average Precision (MAP) (Manning et al., 2008) for comparing the ranked list of semantic features predicted by the model with the flat set of semantic features extracted from WordNet and VerbNet. Average Precision for a ranked list F with respect to a set R of correct features is:

$$\mathrm{AP}_R(F) = \frac{1}{|R|} \sum_{r=1}^{|F|} P(r) \times \mathbf{1}_R(F_r) \qquad (11)$$

where $P(r)$ is precision at rank r and $\mathbf{1}_R$ is the indicator function of set R.

The middle row of Table 4 shows the MAP scores over all the noun or verb usages in the test set, based on four different category sets. As can be seen, the categories induced by our model (ΔH) outperform all the other category sets. The word-type categories are particularly unsuitable for this task, since they provide the least degree of generalization over the semantic properties of a group of words. The categories of Parisien et al. (2008) result in a better performance than word types, but they are still too sparse for this task. However, the average score gained by part of speech tags is also lower than the one by our categories. This suggests that too broad categories are also unsuitable for this task, since they can only provide predictions about the most general semantic properties, such as ENTITY for nouns, and ACTION for verbs. These findings again confirm our hypothesis that a finer-grained set of categories that are extracted directly from the input data provide the highest predictive power in a naturalistic language task such as semantic inference.

6 Grammaticality Judgment

Speakers of a natural language have a general agreement on the *grammaticality* of different sentences. Grammaticality judgment has been viewed as one of the main criteria for measuring how well a language is learned by a human learner. Experimental studies have shown that children as young as five years old can judge the grammaticality of the sentences that they hear, and that both children's and adults' grammaticality judgments are influenced by the distributional properties of words and their context (e.g., Theakston, 2004).

Several methods have been proposed for automatically distinguishing between grammatical and ungrammatical usages (e.g., Wagner et al., 2007). The 'shallow' methods are mainly based on n-gram frequencies of words or categories in a corpus, whereas the 'deep' methods treat a parsing failure as an indication of a grammatical error. Since our focus is on evaluating our category set, we use trigram probabilities as a measure of grammaticality, using Equation 9 with $n = 3$.

As before, we label each test sentence using different category sets, and calculate the probability for each trigram in that sentence. We define the overall grammaticality score of a sentence as the minimum of the probabilities of all the trigrams in that sentence. Note that, unlike the previous tasks, here we do use the content word as a feature in

labeling a test word usage. The actual word form affects the grammaticality of its usage, and this information is available to the human subjects who evaluate the grammaticality of a sentence.

Since we know of no publicly available corpus of ungrammatical sentences, we artificially construct one: for each sentence in our test data set, we randomly move one word to another position.[1] We define the accuracy of this task as the proportion of the test usages for which the model calculates a higher grammaticality score for the original sentence than for its ungrammatical version.

The last row of Table 4 shows the accuracy of the grammaticality judgment task across different category sets. As can be seen, the highest accuracy in choosing the grammatical sentence over the ungrammatical one is achieved by using the PoS categories (0.728), followed by the categories induced by our model (0.715). These levels of accuracy are rather good considering that some of the automatically generated errors are also grammatical (e.g., *there you are* vs. *you are there*, or *can you reach it* vs. *you can reach it*). The results by the other two category sets are lower and very close to each other.

These results suggest that, unlike the semantic inference task, the grammaticality judgment task might require a coarser-grained set of categories which provide a higher level of abstraction. However, taking into account that the PoS categories are manually assigned to the test usages, the difference in their performance might be due to lack of noise in the labeling procedure. We plan to investigate this matter in future by improving our categorization model (as discussed in Section 7). Also, we intend to implement more accurate ways of estimating grammaticality, using an approach similar to that described for word prediction task in Section 4.

7 Discussion

We have proposed an incremental model of lexical category acquisition based on the distributional properties of words. Our model uses an information theoretic clustering algorithm which attempts to optimize the category assignments of the incoming word usages at each point in time. The model can efficiently process the training data, and induce an intuitive set of categories from child-directed speech. However, due to the incremental nature of the clustering algorithm, it does not revise its previous decisions according to the data that it later receives. A potential remedy would be to consider merging the clusters that have recently been updated, in order to allow for recovery from early mistakes the model has made.

We used the categories induced by our model in word prediction, inferring the semantic properties of novel words, and grammaticality judgment. Our experimental results show that the performance in these tasks using our categories is comparable or better than the performance based on the manually assigned part of speech tags in our experimental data. Furthermore, in all these tasks the performance using our categories improves over a previous incremental categorization model (Parisien et al., 2008). However, the model of Parisien employs a number of cluster reorganization techniques which improve the overall quality of the clusters after processing a substantial amount of input data. In future we plan to increase the size of our training data, and perform a more extensive comparison with the model of Parisien et al. (2008).

The promising results of our experiments suggest that an information-theoretic approach is a plausible one for modeling the induction of lexical categories from distributional data. Our results imply that in many language tasks, a fine-grained set of categories which are formed in response to the properties of the input are more appropriate than the coarser-grained part of speech categories. Therefore, the ubiquitous approach of using PoS categories as the gold standard in evaluating unsupervised category induction models needs to be reevaluated. To further investigate this claim, in future we plan to collect experimental data from human subjects performing our suggested tasks, and measure the correlation between their performance and that of our model.

Acknowledgments

We would like to thank Nicolas Stroppa for insightful comments on our paper, and Chris Parisien for sharing the implementation of his model. Grzegorz Chrupała was funded by the BMBF project NL-Search under contract number 01IS08020B. Afra Alishahi was funded by IRTG 715 "Language Technology and Cognitive Systems" provided by the German Research Foundation (DFG).

[1] We used the software of Foster & Andersen (2009).

References

Barbará, D., Li, Y., & Couto, J. (2002). COOLCAT: an entropy-based algorithm for categorical clustering. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 582–589).

Brown, P., Mercer, R., Della Pietra, V., & Lai, J. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Cartwright, T., & Brent, M. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

Clark, A. (2000). Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd workshop on Learning Language in Logic and the 4th conference on Computational Natural Language Learning* (pp. 91–94).

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 59–66).

Fellbaum, C. (Ed.). (1998). *WordNet, an electronic lexical database*. MIT Press.

Fisher, C., Klingler, S., & Song, H. (2006). What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition*, *101*(1), 19–29.

Foster, J., & Andersen, Ø. (2009). GenERRate: generating errors for use in grammatical error detection. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications* (pp. 82–90).

Frank, S., Goldwater, S., & Keller, F. (2009). Evaluating models of syntactic category acquisition without using a gold standard. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, 1535–1540.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (Vol. 45, p. 744).

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

Kemp, N., Lieven, E., & Tomasello, M. (2005). Young Children's Knowledge of the" Determiner" and" Adjective" Categories. *Journal of Speech, Language and Hearing Research*, *48*(3), 592–609.

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Harvard University Press Cambridge, Mass.

Lesher, G., Moulton, B., Higginbotham, D., & Alsofrom, B. (2002). Limits of human word prediction performance. *Proceedings of the CSUN 2002*.

Li, T., Ma, S., & Ogihara, M. (2004). Entropy-based criterion in categorical clustering. In *Proceedings of the 21st International Conference on Machine Learning* (p. 68).

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates Inc, US.

Manning, C., Raghavan, P., & Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press New York, NY, USA.

Meila, M. (2003). Comparing Clusterings by the Variation of Information. In *Learning theory and kernel machines* (pp. 173–187). Springer.

Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*(5), 678–686.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Naigles, L., & Hoff-Ginsberg, E. (1995). Input to Verb Learning: Evidence for the Plausibility of Syntactic Bootstrapping. *Developmental Psychology*, *31*(5), 827–37.

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental bayesian model for learning syntactic categories. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning.*

Redington, M., Crater, N., & Finch, S. (1998). Distributional information: A powerful cue for ac-

quiring syntactic categories. *Cognitive Science: A Multidisciplinary Journal*, *22*(4), 425–469.

Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Unpublished doctoral dissertation, University of Pennsylvania.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Theakston, A.(2004). The role of entrenchment in childrens and adults performance on grammaticality judgment tasks. *Cognitive Development*, *19*(1), 15–34.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*(01), 127–152.

Wagner, J., Foster, J., & van Genabith, J.(2007). A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. *Proceedings of EMNLP-CoNLL-2007*.

Zhai, C., & Lafferty, J.(2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, *22*(2), 214.