Lexical Category Acquisition as an Incremental Process

Afra Alishahi
afra.alishahi@coli.uni-sb.de
Computational Linguistics and Phonetics
Saarland University, Germany

Grzegorz Chrupała
gchrupala@lsv.uni-saarland.de
Spoken Language Systems
Saarland University, Germany

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 the 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 shown to be a reliable cue for the formation of the lexical categories (Redington et al., 1998; Mintz, 2003). Several computational models have used distributional information for categorizing words (e.g. Brown et al., 1992; Schütze, 1993; Redington et al., 1998; Clark, 2000; Mintz, 2002). The majority of these models use iterative, unsupervised methods that 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 hint at distributional cues as a useful source of information for this purpose.

The process of learning word 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 must be developed to investigate whether the distributional information is equally powerful in an online process of word categorization. There have only been a few previous attempts at applying an incremental method to category acquisition. 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. The model proposed by Parisien et al. (2008) uses a Bayesian clustering algorithm that can cope with ambiguity, and shows the developmental trends observed in children (e.g. the order of acquisition of different categories). However, their fully Bayesian implementation is computationally expensive. Moreover, when measuring the similarity between two contexts, the model is sensitive to mismatches between any pair of context features, which results in the creation of sparse clusters. To overcome the problem, they introduce a bootstrapping mechanism which improves the performance, but adds substantially to the computational load.

We propose an efficient incremental model for clustering words into categories based on their local context. Each word of a sentence is processed and categorized individually based on the similarity of its content (the word itself) and its context (the surrounding words) to the existing clusters. We test our model on a corpus of child-directed speech from CHILDES (MacWhinney, 2000). Over time, the model learns a fine-grained set of word categories that are intuitive and can be used in a variety of tasks. We evaluate our model on a word prediction task, where a missing word is guessed based on its context. We also use our model to infer the semantic properties of a novel word based on the context it appears in. In both tasks, we show that our induced categories outperform the part of speech tags used for annotating the corpus.

An Incremental Category Acquisition Model

We propose an online clustering algorithm for categorizing word usages (i.e. tokens) in unannotated text, inspired by online spherical K-means (Zhong, 2005). The algorithm categorizes the word usages one at a time, and updates the existing categories or forms new ones as a result. For each word usage, a new category C_{new} is created. A similarity score is then measured between C_{new} and each of the existing categories. If the similarity between C_{new} and the most similar category is higher than a certain threshold θ_w, the two categories are merged. Since the categories are formed incrementally and as a response to the order of input usages, the model may create unnecessary categories at the beginning: if two words that have the same syntactic properties appear in two different contexts early on, they might be put into two different categories. Therefore, we propose a revision mechanism to recover from such mistakes: once a new category C_{new} is merged with an existing one, it is again compared with the existing categories and merged with the closest one if their similarity exceeds a second threshold parameter θ_c. The algorithm is summarized in Algorithm 1.

Following Redington et al. (1998) and Mintz (2003), we estimate the similarity of two categories based on the content feature (the target word), and the context features (two preceding and two following words). Each category is represented as a vector which is the mean of the feature vectors corresponding to all the word usages that were added to that category at some point in learning. The mean vector of a category is immediately updated when it is merged with another one. We use the dot product of the feature vectors representing two categories as our similarity metric.

Algorithm 1 Incremental Word Clustering

For every word usage w:

- Create new cluster C_{new}
- Add $\Phi(w)$ to C_{new}
- $C_w = \text{argmax}_{C \in \text{Clusters}}$ **Similarity**(C_{new}, C)
- If **Similarity**$(C_{new}, C_w) \geq \theta_w$
  - merge C_w and C_{new}
  - $C_{next} = \text{argmax}_{C \in \text{Clusters} - \{C_w\}}$ **Similarity**(C_w, C)
  - If **Similarity**$(C_w, C_{next}) \geq \theta_c$
    - merge C_w and C_{next}

where **Similarity**$(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ and the vector $\Phi(w)$ represents the context features of the current word usage w.

Evaluation

Many unsupervised models of lexical category acquisition treat the traditional part of speech (PoS) tags as the gold standard, and evaluate their induced categories by how closely they resemble the high-level PoS categories such as noun, verb and adjective (e.g. Parisien et al., 2008). However, it is not at all clear whether humans form the same types of categories. In fact, many language tasks seem to rely on finer-grained classes (e.g. animates, food or motion verbs).

We evaluate the categories formed by our model through two different tasks. In the first task, we use the context to predict the target word. In the second task, the same context is used to infer the semantic properties of a novel word. We use a corpus of child-directed speech, and show that the induced categories outperform the PoS tags used for manually annotating the same corpus.

Experimental Setup

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 used about 3300 word usages for one child (Anne) as development data, based on which we empirically set the parameters $\theta_w = 2^7 \times 10^{-3}$ and $\theta_c = 2^{10} \times 10^{-3}$. We used half of the Anne conversations as the training set, and a small portion of Becky's conversations as the test set. We discarded all one-word sentences from the test set, as they do not have the context necessary for our evaluation tasks. Table 1 gives more details on the datasets used.

In both tasks described below, we trained the model on our training set, which resulted in a set of 690 categories. Table 2 shows some of the categories learned from the training set. We then froze the categories, and used them to label the word usages in the test set. However, we did not use the content feature for categorizing the test words, since the tasks involve the prediction of the target word or its properties.

Table 1: Experimental data

Data Set	Corpus	#Sentences	#Words
Development	Anne	857	3,318
Training	Anne	19,300	78,000
Test	Becky	1,560	5,500

Table 2: Example clusters

Most frequent features for the focus word
do, are, will, have, can, has, does, had, were, could, …
train, cover, one, tunnel, hole, king, door, fire-engine, …
's, is, was, in, then, goes, on, …

Most frequent features for the previous word
bit, little, good, big, very, long, few, drink, funny, …
the, a, this, that, her, there, their, our, another, enough, …
're, 've, want, got, see, were, do, find, going, know, 'll, …

Predicting a Word based on the Context

Humans can predict a word based on the context it is used in with remarkable accuracy (e.g. Lesher et al., 2002). We simulate this behavior, where a missing word is guessed based on its context. For each categorized word usage in the test set, we predict the target word based on its labeled category: the ranked list of word forms corresponding to the content feature of the category represent this prediction. We compute the reciprocal of the rank of the target word in this list. Table 3 shows the average reciprocal rank for the 5500 words in the test set.

To compare our categories with the standard PoS labels, we used the annotated version of our training set to form a similar feature representation for the PoS categories: all the word usages that were labeled with the same tag were grouped together, and their contexts were used to calculate the mean feature vector for each tag. We applied the same word prediction method on the test set using the PoS categories, and calculated the reciprocal rank. The average score over all word usages in the test set is shown in Table 3. As can be seen, the average reciprocal rank based on the induced categories is almost three times higher than the one based on the PoS categories ($p < 10^{-16}$, paired t-test). The results suggest that a larger set of categories which embodies finer-grained distinctions is more apt for a word prediction task.

Inferring Semantic Propeties of a Novel Word

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). To study a similar effect in our model, we associate

Table 3: Results for the evaluation tasks, based on two sets of categories

Word Prediction

Category type	Mean recip. rank
PoS	0.078
Induced	**0.231**

Semantic induction

Category type	Avg. dot product
PoS	0.031
Induced	**0.048**

each word with a representation of its semantic properties. Following Fazly et al. (2008), we extract a semantic feature vector for each word from WordNet. These features are not used in clustering; rather, to each category we associate a semantic feature vector which is the mean of the semantic vectors of all the words that at some point have been added to that category. However, we limit our evaluation to nouns and verbs, since WordNet is mainly developed based on these two categories.

Similar to the word prediction task, we treat the semantic features of the category assigned to a novel word as the prediction of the model for the semantic properties of that word. We compare the semantic features of the category with the semantic features of the target word, using the dot product of the two vectors. Similarly, we build a semantic feature vector for the PoS categories based on the training set, and compare the semantic vector of each labeled noun or verb usage in the test set with the semantic vector of the corresponding PoS category.

Table 3 shows the average dot product for the test set, based on both the categories induced by our model and the PoS categories. The average measure based on our categories is more than 1.5 times larger than the one based on the PoS categories ($p < 10^{-16}$, paired t-test), suggesting that the predicted semantic properties based on our induced categories are a much better match for the actual properties of the target word. These results again confirm that a finer set of categories are more useful in inferring the semantic properties of an unknown word based on its context.

Discussion

We have proposed an incremental model of lexical category acquisition based on distributional properties of words, using an efficient clustering algorithm. Our model induces an intuitive set of categories from child-directed speech, and can use them in word prediction and the inference of the semantic properties of a word from context. We argue that for these tasks, a finer-grained set of categories such as the ones developed by our model is more appropriate than the traditional coarse-grained categories used for corpus annotation.

In future, we plan to use the predicted categories of the previous words as additional features, and investigate their impact on the categories. Further, we intend to use the categories in other tasks such as lexical disambiguation, and compare the behavior of the model to human performance.

References

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-volume 7* (pp. 91–94).

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proceedings of the 30th annual conference 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.

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*.

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

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

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st annual meeting on association for computational linguistics* (pp. 251–258).

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.

Zhong, S. (2005). Efficient online spherical k-means clustering. In *2005 ieee international joint conference on neural networks, 2005. ijcnn'05. proceedings* (Vol. 5).