Robust Induction of Parts-of-Speech in Child-Directed Language by Co-Clustering of Words and Contexts

Richard E. Leibbrandt
School of Computer Science, Engineering and Mathematics
Flinders University

richard.leibbrandt@flinders.edu.au

David M W Powers
School of Computer Science, Engineering and Mathematics
Flinders University

david.powers@flinders.edu.au

Abstract

We introduce Conflict-Driven Co-Clustering, a novel algorithm for data co-clustering, and apply it to the problem of inducing parts-of-speech in a corpus of child-directed spoken English. Co-clustering is preferable to unidimensional clustering as it takes into account both item and context ambiguity. We show that the categorization performance of the algorithm is comparable with the co-clustering algorithm of Leibbrandt and Powers (2008), but out-performs that algorithm in robustly pruning less-useful clusters and merging them into categories strongly corresponding to the three main open classes of English.

1 Introduction

The problem of unsupervised part-of-speech induction has received considerable attention in computational linguistics (for a recent comparison of several influential models, see Christodoulopoulos, Goldwater & Steedman, 2010). A common approach is to estimate the parameters of a generative model given the natural language data, with the model usually a variant of a Hidden Markov Model (e.g. Goldwater & Griffiths, 2007; Berg-Kirkpatrick, Côté, De Nero & Klein, 2010; Moon, Erk & Baldridge, 2010). These models are often evaluated on corpora of formal, written English, such as the Penn Treebank, rather than on natural, spoken language, and typically the aim of these studies is to improve the state-of-the-art of POS induction using various techniques from machine learning, with an implicit focus on devising techniques that can be used in practical applications.

In the current paper, on the other hand, our focus is on part-of-speech induction mechanisms that children might use when learning their first language. Hence, we are interested in models that are motivated by psychological considerations, rather than by a more abstract mathematical or statistical grounding. In language acquisition research, a typical approach to part-of-speech induction is to make use of clustering. We will review this work and argue for the particular utility of two-mode clustering or co-clustering approaches, before presenting two novel co-clustering techniques and evaluating their performance in part-of-speech tagging on a corpus of child-directed English.

1.1 Clustering and co-clustering approaches to part-of-speech induction in language acquisition research

Single-mode clustering approaches

Clustering algorithms operate on a two-dimensional matrix where the rows and columns in this context represent words and the linguistic contexts in which they appear, taken from a corpus of natural language, and the cells of the matrix contain frequency counts of how often a word occurs in a particular context. It has often been proposed that children might make use of information about the contextual distribution of usage of words to induce the parts-of-speech of their native language (e.g. Maratsos & Chalkley, 1980), and work by, e.g., Redington, Chater & Finch (1998) and Clark (2000), showed that parts-of-speech can indeed be induced by

clustering together words that are used in similar contexts in a corpus. Clustering word types together does not take into account the fact that the part-of-speech of a word type may change depending on the context in which it is used. One of the most influential models in part-of-speech induction in language acquisition, the Frequent Frames model of Mintz (2003), addresses this issue by forming clusters of the contextual frames in which words are used, rather than the words themselves. The idea is that the contexts define the part-of-speech, rather than the words themselves. This model attains high, but not perfect results in part-of-speech tagging for English child-directed speech; part of the reason is that even frames are sometimes ambiguous in the parts-of-speech that they can accommodate, and Erkelens (2008) has shown that this problem is more pronounced when the Frequent Frames approach is applied to Dutch material. In general, however the set of frame contexts is chosen, the problem of contextual ambiguity is likely to present itself. Hence, an approach is needed in which both words and contexts can be associated with multiple categories. Techniques of co-clustering, also called biclustering or two-mode clustering, (see Madeira & Oliveira, 2004, Van Mechelen et al., 2004, for reviews), represent one such approach.

Co-clustering approaches

Single-mode clustering forms clusters of elements in one dimension of the matrix (either rows or columns) by grouping together elements on the basis of similar co-occurrence with elements of the other dimension. Co-clustering techniques, on the other hand, form clusters on the basis of similarity between rows and similarity between columns simultaneously. Co-clustering is therefore able to assign row and column elements to the same clusters. We can distinguish between row-column clustering methods which assign each row and each column to a particular cluster, and data clustering methods which assign each individual non-empty cell of the matrix to a cluster. Some co-clustering methods allow for overlapping clusters, i.e. in row-column methods by allowing rows and columns to belong to more than one cluster, or in data clustering methods by allowing cells in the matrix to belong to more than one cluster. Co-clustering algorithms have been shown to be useful in many applications, notably in the analysis of gene expression data (Madeira & Oliveira, 2004).

There are good reasons to prefer a co-clustering approach over a single-mode categorization approach in part-of-speech induction. In natural language, including child-directed speech, there are many cases where a word appears in a context that does not specify the part-of-speech exactly, but allows several possibilities, while at the same time, the word is also ambiguous in its part-of-speech. Co-clustering is able to deal with part-of-speech ambiguity at the level of word and frame simultaneously. For example, a common frame in child-directed speech in English is "That's X.", where the word that fills the X slot could be a noun ("That's ice-cream.") or an adjective ("That's pretty."). Simultaneously, the word "mean" can be used as either a verb or an adjective (the nominal usage is rare in child-directed speech). A single-mode clustering algorithm that aims to assign a part-of-speech to the word "mean" in "That's mean" will be unable to decide between the allowed parts-of-speech for the frame, if frames were clustered, and between the allowed parts-of-speech for the words, if words were clustered. However, a co-clustering approach that assigned "That's X" to both the categories noun and adjective, and "mean" to the categories verb and adjective, would be able to deduce that the only category that the word and the frame have in common is adjective, and therefore that this is the correct category. In this way, co-clustering is better able to deal with linguistic ambiguity.

Even apart from its practical utility in part-of-speech induction, co-clustering is broadly compatible with a psychological outlook that conceives of part-of-speech development in terms of associative learning (see e.g. Shanks, 1995). Under this view, parts-of-speech are mental categories that are formed by repeated exposure to words used in context, in combination with whatever semantic construal the language-learning child places on the utterances she hears.

Only a few studies have applied co-clustering to part-of-speech induction with child-directed language (but see Freitag, 2004, for part-of-speech induction with co-clustering on adult-directed language in the Penn Treebank). The pioneering work in this regard was the EMILE system of Adriaans and colleagues (Adriaans, 1992), which formed co-clusters of word-context combinations as a step in the process of inducing

rules for a categorial grammar. While the grammars formed in this way perform well, EMILE typically produces large, overlapping categories which do not correspond to the parts-of-speech of English (Adriaans, 1999). Hence, it is difficult to evaluate the accuracy of EMILE's part-of-speech tagging against a gold standard.

Leibbrandt & Powers (2008) applied co-clustering to a corpus of English child-directed speech, yielding accuracy comparable to that obtained by the Frequent Frames model of Mintz (2003). This approach was also able to outperform Frequent Frames in tagging child-directed data in Dutch (Leibbrandt & Powers, 2010).

In this paper, we extend the work of Leibbrandt & Powers (2008, 2010) by describing and evaluating a novel co-clustering technique for part-of-speech induction. In Section 2 we present the Conflict-Driven Co-Clustering algorithm, and in Section 3 we evaluate its performance in part-of-speech tagging of a corpus of child-directed speech. We show that the algorithm delivers performance comparable to that of both the Frequent Frames model of Mintz (2003) and the co-clustering work by Leibbrandt & Powers (2008, 2010), and is more robust than the earlier work in automatically discovering the main English open classes of noun, verb and adjective, discarding smaller and less-easily interpretable categories. In Section 4 we consider reasons for these results and point to future directions for this work.

2 Conflict-Driven Co-Clustering

The Conflict-Driven Co-Clustering (CDCC) algorithm is a row-column-based co-clustering algorithm. It creates an initial clustering of words into a set of clusters, and a simultaneous clustering of frames into the same set of clusters. Only a few word and frame types are clustered to start with, and hence this initial clustering is inadequate to account for the empirical co-occurrence data (as explained below). From this starting point, the CDCC algorithm iteratively adds frames and clusters to the clusters, until all of the co-occurrence data is accounted for.

We make the assumption that there exist a number of parts-of-speech in the target language, and that a particular word used in a particular frame context belongs to only one part-of-speech[1]. We also assume that the word type is a cue to the part-of-speech, and that the same is true of the frame type. Finally, each word type and frame type is presumed to have the potential to be associated with more than one part-of-speech.

Suppose, then, that we (in this case, the co-clustering algorithm, but also, potentially, a child learning the target language) already have some notion of the parts-of-speech to which a particular frame type *f* "belongs", and the parts-of-speech to which a word type *w* belongs. Then when we encounter an instance (i.e. a token) of the word type *w* used in the context of the frame type *f*, and wish to assign a part-of-speech to this instance, the only viable candidates (based on our knowledge at the time) are those parts-of-speech that both *w* and *f* have in common. Should there be multiple such candidates, a part-of-speech tagging algorithm might resort to combining information about the probabilities of *f* and *w* belonging to each candidate in order to select a "winner". However, when there is *no* such candidate (word and frame have no part-of-speech in common), this presents a problem for part-of-speech tagging. Such a situation is an instance of the "conflicts" from which CDCC derives its name.

More concretely, we can represent the cluster membership of each of the J words under consideration as a $J \times K$ matrix W, where K is the number of clusters, and $W_{jk} = 1$ if word j is a member of cluster k, and 0 otherwise. Similarly, the cluster membership of each of the I frames is represented by the $I \times K$ matrix F, where $F_{ik} = 1$ if frame i belongs to cluster k, and 0 otherwise.

The $I \times J$ matrix D represents the co-occurrence data obtained from the corpus, where $D_{ij} = 1$ if word j occurs in the context of frame i in the corpus, and 0 otherwise. Then a conflict exists whenever $D_{ij} = 1$ and the dot-product $W_j \cdot F_i = 0$.

We can think of the possibilities described by the cluster membership matrices W and F as accounting for the word-frame co-occurrences described in D: if a word and frame can occur together, there must be at least one part-of-speech to which they both belong. Conflicts occur where cells in the D matrix are not yet accounted for in this way. The problem to be solved in this case, therefore, is to remove all

[1] There are examples, even in the corpus used in this experiment, for which this assumption does not seem to hold; however, these examples are relatively infrequent enough to warrant its use as a useful heuristic.

instances of conflict. Because the *D* matrix is empirically given, the only way to remove conflict is to modify the *F* and *W* matrices so that all co-occurrences in *D* can be accounted for.

Figure 1 illustrates some cases of conflict and resolved conflict between word and frame. Initially, the utterance "Shall I brush it?" contains a conflict, because the frame "Shall I X it?" is allocated to the Verb category, but "brush" is not yet allocated to any category. The conflict might be resolved by adding "brush" to the Verb category. Later, when we consider the utterance "There's your brush", a conflict would occur if "brush" was allocated to Verb only and "There's your X" was allocated to Noun only. Suppose that the conflict was resolved correctly by also adding "brush" to the category Noun (in addition to already being allocated to Verb). Then when the utterance "Don't brush it" is encountered, there is no conflict, as both "Don't X it" and "brush" are allocated to the Verb cluster, and hence the allocations are compatible.

Shall I brush it?	N	V	A
brush	0	**0**	0
Shall I X it?	0	**1**	0

There's your brush.	N	V	A
brush	**0**	**1**	0
There's your X.	**1**	**0**	0

Don't brush it.	N	V	A
brush	1	1	0
Don't X it.	0	1	0

Figure 1. Three instances of conflict and non-conflict. In the top example, *brush* and *Shall I X it?* are in conflict, in the middle example, *brush* and *There's your X* are in conflict, and in the lower example there is no conflict. (N = Noun, V = Verb, A = Adjective)

An open problem is then how best to calculate the cluster membership matrices *W* and *F* so as to remove all conflicts. One obvious "solution" would be to simply add membership of every cluster to every word and frame. While this would remove all conflicts, it is clearly not a useful basis for part-of-speech tagging, and violates our sense that not every word or context can belong to every part-of-speech.

A better approach might be to start with a very sparse pair of initial matrices for *W* and *F*, which greatly under-determine the co-occurrence matrix *D*, and then add cluster memberships to individual frames and words (changing 0s to 1s in *F* and *W*) if adding them would help to solve conflicts.

We still need to decide which cluster memberships to add, and a useful principle might be to add memberships parsimoniously, i.e. to try to minimize the number of new memberships added to *F* and *W*. The CDCC algorithm takes a greedy approach to this problem. On each iteration, it simply adds the single cluster membership (word or frame) that would resolve the largest number of conflicts existing at that time. The set of remaining conflicts is then recalculated, and the cluster membership that again resolves the greatest number of conflicts is added, with the process being repeated until all conflicts have been resolved.

The only remaining point to specify is how the algorithm gets started, i.e. how the *W* and *F* matrices are initialized. It would be desirable to begin with just a small number of "ground truths", i.e. a small number of category memberships, for only a few frames and words, that are well-established in advance. The rest of the values in the membership matrices are then bootstrapped from this starting point by referring to the co-occurrence matrix.

The initial values with which *W* and *F* are "seeded" can come from any source: for instance, they may be the result of a process of semantic category formation (e.g. Macnamara, 1982; Pinker, 1984), so that words that refer to physical objects are flagged as belonging to one category, and words for actions marked as belonging to another category (bear in mind that this does not preclude these words from later also being assigned to other categories). This process might also be extended to frames that reliably contain words referring to objects, actions, physical properties, etc. In computational work on language acquisition, proxies for these categories might be obtained from lists of early-acquired words, possibly in combination with norms on word imageability. In less acquisition-oriented work, seeds may be obtained from manually annotated examples, so that this becomes a semi-supervised approach to part-of-speech tagging.

In the experiment reported here, we decided to obtain our seed information entirely from the same word-frame co-occurrence matrix *D* used later to expand the *W* and *F* matrices, and we did so along the same lines as followed by Leibbrandt & Powers (2008, 2010). Consequently, our results are prone to some of the shortcomings of the earlier work, as

discussed later. We emphasize that the choice of seeding algorithm is not part of the CDCC algorithm proper, and informal experimentation has shown that the performance of CDCC is highly dependent on the accuracy of the initial seed information.

2.1 CDCC Algorithm

The conflict-driven co-clustering algorithm (pseudo-code is presented in Box 2) attempts to find a conflict-free allocation of categories to words and frames. It does so by repeatedly removing the largest existing conflict until no conflicts remain.

In what follows, we use the term "co-item" to refer to those items with which an item (word or frame) co-occurs in *D*, i.e. the co-items of a word type are the frame types in which it has occurred, and the co-items of a frame type are the word types that have occurred in it. Conflicts between items and their co-items are removed by simply allocating those additional categories to items that they would need in order to no longer be in conflict with the co-items. Conflicts are not resolved in random order; instead, the conflict resolution option that would resolve the largest number of conflicts is chosen at every step. In this way, the membership vectors for each of the words and frames are adjusted so as to converge onto the "correct" allocation. When no more changes can be made to the membership vectors, the algorithm halts.

The algorithm works in batch mode, considering the entire data matrix at once. For every item (whether word or frame), the set of co-items that are currently in conflict with the item is collected. Using the current membership matrices *W* and *F*, the algorithm allows each co-item to cast one vote for every category to which it is currently allocated (i.e. co-items cast votes to have particular categories added to the item's allocations). Per definition, these are categories that the target item does not have in its membership vector, so that adding that category to the item's membership vector would resolve the conflict between the item and that particular co-item; however, the point of voting is to find the single change that would result in the *largest number* of conflict resolutions at once. The number of votes for each category is determined in this way for every target item (every word and every frame). The suggested category allocation that has received the largest number of votes over all words and all frames is designated the "winner", and the category in question is added to the membership vector of the item in question.

CDCC:

D: co-occurrence matrix of frames and words

F, W: membership matrices describing the categories to which each of the frames and words may belong. F and W are initialized prior to running CDCC, for instance using unsupervised clustering as in Box 2. F[k][i] = 1 if frame i is able to belong to cluster k, and 0 otherwise, and similarly for W.

```
repeat until convergence (see text)
  for i = 1 to I
    for j = 1 to J
      if D[ i ][ j ] = 1
          conflict = true
          for k = 1 to K
              if (F [ k ][ i ] = 1
              and W [ k ][ j ] = 1)
                  conflict =false
          if conflict
              tallyVotes(i, j)
  find k1 such that FrameVotes[ k1 ][ i ] =
      max cell in FrameVotes
  find k2 such that WordVotes[ k2 ][ j ] =
      max cell in WordVotes
  if FrameVotes[k1][ i ] > WordVotes[k2][ j ]
      F [ k1 ][ i ] = 1
  else
      W [ k2 ][ j ] = 1

tallyVotes(i, j):
  for k = 1 to K
    if (F[ k ][ i ] = 0 and W [ k ][ j ] = 1)
          FrameVotes[ k ][ i ] += 1
    else if (F [ k ][ i ] = 1 and W [ k ][ j ] = 0)
          WordVotes[ k ][ j ] += 1
```

Box 1. Conflict-Driven Co-Clustering Algorithm.

One of the benefits of the voting system is that it is self-correcting. If an item which is, say, a Noun, is incorrectly not assigned to the cluster corresponding to Nouns, then it will cast one incorrect vote each time to change the allocation of each of its co-items. However, the co-items are likely to be Nouns in most cases, and hence

to occur in other Noun frames, which will in most cases lend them the Noun allocation, so that they will vote en masse to change the allocation of the incorrectly allocated item to Noun.

The product of the CDCC algorithm is a fairly conservative allocation of (potentially multiple) clusters to each of the words and frames.

3 Evaluation of the algorithms

The CDCC algorithm was applied to a corpus of child-directed speech, after which individual tokens of word-frame co-occurrences were categorized into one of the co-clusters produced by the algorithm, as described below.

3.1 Data Set

The data set used was the same as in Leibbrandt & Powers (2008), namely the child-directed portion of the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001) obtained from the CHILDES project (MacWhinney, 2000). This corpus is supplied with a manual part-of-speech tagging, which was used as the 'gold standard' correct tagging against which the categorization produced by CDCC was evaluated.

3.2 Extraction of Contextual Frames

Contextual frames were extracted from the corpus following the method in Leibbrandt & Powers (2008). Frames were formed from utterances in the corpus by replacing all but the most frequently-occurring words in the corpus with a placeholder symbol, turning corpus utterances into lexically-based schematic template sentences with slots that can be filled by inserting single words (for example, "Don't X it", "That's your X", "It's very X"). Frequency counts were collected of the number of occurrences of each word in each of the contextual frames, and the resulting data matrix was filtered to contain only those elements that attained a certain level of support, i.e. frames that occurred with 5 or more distinct word types, and words that occurred in 5 or more frame types. The resulting data matrix was used to obtain seed category membership information for selected words and frames, as described in the next section.

3.3 Seed Information

The first step in obtaining "ground truth" seed information for running the CDCC algorithm (pseudocode shown in Box 2) is to perform a standard one-mode clustering of the (L2-normalized) frame vectors of the co-occurrence matrix *D*, producing clusters of contextual

D: co-occurrence matrix, such that D [i][j] = 1 if word j has co-occurred with frame I, 0 otherwise.

Allocation: Cluster membership vector for frames, obtained from hard clustering algorithm, such that Allocation[i] = k if frame *i* is allocated to cluster *k*.

```
Initialize ClusterCoocc[ K ][ J ] to all zeroes.
for i = 1 to I
  for j = 1 to J
    if D[ i ][ j ]
      ClusterCoocc [ Allocation[ i ] ] [ j ] += 1
for k = 1 to K
  sum = sum(ClusterCoocc [ k ])
  for j = 1 to J
    Distribution[ k ][ j ].index  = j
    Distribution[ k ][ j ].value =
      ClusterCoocc [ k ][ j ] / sum
  Sort Distribution[ k ] by value (descending)
  cumulativeProportion = 0;  j = 0
  repeat until cumulativeProportion ≥ η
    j += 1
    index = Distribution[ k ][ j ].index
    value = Distribution[ k ][ j ].value
    SeedWords[ k ] [index]  = 1
    cumulativeProportion += value
for each pair (SeedWords[a], SeedWords[b]),
  a ≠ b
    Remove all words that occur in both
    SeedWords[a] and SeedWords[b]
for i = 1 to I
  for j = 1 to J
    if D [ i ][ j ]
      for k = 1 to K
        if SeedWords[ k ][ j ] = 1
          SeedFrames[ k ][ i ] = 1
for each pair (SeedFrames[a], SeedFrames[b]),
  a ≠ b
    Remove all frames that occur in both
    SeedFrames[a] and SeedFrames[b]
```

Box 2. Seed frame and word selection algorithm.

frames (hierarchical clustering with average linkage was used in this experiment).

Next, we select sets of words that are particularly distinctive of each of the frame clusters. The assumption is that words that occur in a large number of frame types from a particular cluster are good representatives of that cluster. Hence, for each cluster, words are ranked in order of the number of distinct frame types from the cluster in which each word has occurred, and are added one-by-one to the seed-word set for the cluster, until the cumulative proportion of total distinct-frame counts accounted for exceeds a threshold (set to 0.25 in this experiment). Once all seed-word sets have been collected in this way, seed-words which occur in the sets of more than one cluster are discarded.

Next, a seed-frame set is created for each cluster, consisting of all frames which occurred with seed-words from that cluster and did not occur with a seed-word from any other cluster. The resulting seed sets are arguably the words and frames that are the most distinctly associated with each cluster. The process described above can be considered to produce similar results to a psychological process of association between clusters and words, where the strength of association between the cluster and the word is strengthened each time the word is used in a frame that is strongly associated with that cluster already. Each distinct frame is considered to contribute an equal amount of activation strength to the word, regardless of its own frequency of occurrence in the input, so that this association process is sensitive to the type frequency of frames co-occurring with the word in question, rather than to the token frequency. A wider range of co-occurring frames constitutes more robust evidence that the word does indeed belong with the cluster (and most likely possesses many of the semantic attributes that are associated with the cluster). For evidence that the type frequency of words occurring in a frame aids generalization, see Bybee (1985, 2006).

The algorithm maintains a binary-valued allocation vector for each frame and each word of length K, where K is the number of clusters. The k'th value in the allocation vector is 1 if the word or frame can belong to cluster k, and 0 if not. In this way, the algorithms deal with the ambiguity of both words and frames, by allowing an item to belong to more than one cluster. For every cluster k, the k'th value of the allocation vector of every seed word and every seed frame of cluster k is initialized to 1, and all other values are set to 0.

3.4 Categorization

For the purpose of evaluation, we categorize each of the instances of word-frame co-occurrences in the data matrix *D* by combining the word and frame cluster information contained in the membership matrices *W* and *F*. When classifying a particular instance of word *w* used in frame *f*, if there exists a unique a cluster *c* such that *w* and *f* have both been allocated to *c* (in a majority of cases in this experiment, there was such a unique cluster), then the word-frame combination is classified as belonging to the cluster in question. In cases where the word and frame have more than one cluster in common, we fall back on estimating the amount of evidence that the word and frame separately belong to each of the clusters. The fallback values for each word and frame are calculated as the proportion of co-items of the word or frame that are allocated to each cluster. The fallback value of the word is multiplied by the fallback value of the frame, for each cluster separately, and the cluster with the highest product is selected as the category to which the frame-word combination is assigned.

3.5 Evaluation Measures

Results are reported in terms of standard measures of precision, recall and F-score, with random baselines in parentheses. These measures were calculated, as is customary in unsupervised categorization, by a pair counting approach that constructs a confusion matrix based on whether *pairs of elements* are assigned to the same category in the gold-standard, and also in the clustering model (see e.g. Mintz, Newport & Bever, 2002). Because of several well-known shortcomings of precision and recall (e.g. Powers, 2003; Rosenberg & Hirschberg, 2007), we also report the Informedness measure (Powers, 2003), which corresponds to the probability that the predictions made by the algorithm are informed, in the sense of making correct use of information.

For a 2×2 contingency table with the symbols *a*, *b*, *c* and *d* respectively indicating the number of true positives, false positives, false negatives and true negatives, Informedness is given by

$$I = \frac{a}{a+c} - \frac{b}{b+d}.$$

Informedness can thus be expressed as Recall for a particular cluster, discounted by the proportion of all non-category items that occur in that cluster. Informedness is equivalent to the well-known delta-P formula expressing association strength in human associative learning (e.g. Shanks, 1995). For a supervised classification problem, with a table of arbitrary dimensions *m*×*m*, Informedness is calculated for the 2×2 contingency table of each category in turn, and the Informedness values for all categories are combined in a weighted sum, where the weight for each category is the proportion of word tokens assigned to that category by the algorithm (i.e. the algorithm's bias to assign instances to the category). In unsupervised cases, it is not obvious how to associate clusters with gold-standard categories. In this case, weighted Informedness values are calculated for every possible 1-to-1 mapping between gold standard categories and clusters, and the highest of these Informedness values is selected.

For evaluation, we made use of only those tokens that were assigned to one of the three major open-class categories (nouns, verbs and adjectives).

	HC	**CDCC**	**LP08**	**FreqF**
Precision	0.844 *(0.559)*	0.888 *(0.559)*	0.900 *(0.559)*	0.90
Recall	0.774 *(0.513)*	0.911 *(0.574)*	0.886 *(0.551)*	0.91
F	0.808 *(0.535)*	0.899 *(0.566)*	0.893 *(0.555)*	0.90
I	0.708	0.800	0.814	n/a

Table 1. Performance of clustering-based part-of-speech induction methods. Random baseline values in italics. Baseline value for Informedness is zero. *HC* = Hierarchical Clustering (one-dimensional); *CDCC* = Conflict-Driven Co-Clustering; *LP08* = replication of Leibbrandt & Powers (2008); *FreqF* = Frequent Frames (results from Mintz, 2006, baseline and Informedness scores unknown).

3.6 Results

The results of categorization according to the *CDCC* algorithm is shown in Table 1. For comparison, we have also shown the results of categorization with three other algorithms, namely: *LP08*, a replication of Leibbrandt & Powers (2008); *FreqF*, the results from Mintz (2003) for the Frequent Frames model applied to the same corpus as used here; and *HC*, the results from categorizing a word-frame combination according to the cluster of the frame only, where the frame clusters are the ones derived in the one-way clustering step that produced the seed information for CDCC.

The results show that CDCC is competitive in its categorization performance with both the LP08 and FreqF approaches. Comparing Informedness and F-scores against their random baselines, the performance of LP08 is only slightly better than that of the two new algorithms (random baseline values were not reported by Mintz, 2003). Importantly, CDCC (as well as LP08) performs much better than the hard clustering HC from which it derives its seed information, showing that co-clustering improves categorization.

3.7 Robustness of induced parts-of-speech

We have not yet said much about the number of clusters formed by the co-clustering algorithms. This number could conceivably be influenced by the number of clusters formed by the initial one-way clustering algorithm, which is often (as it was in our experiment) a parameter under control of the experimenter. However, the number of parts-of-speech produced by a part-of-speech induction algorithm should be relatively immune to manipulations of algorithmic parameters. A related issue is that the parts-of-speech produced by clustering approaches are often unsatisfactory from a linguistic point of view, as they don't correspond exactly to the expected parts-of-speech of the target language (see also Schütze, 1995). We regard it as desirable for a part-of-speech induction method to account for at least the main open-class parts-of-speech of English (nouns, verbs, adjectives and adverbs), and to be able to produce these without undue coercion.

Therefore, it is of interest to consider how the number of parts-of-speech produced by the co-clustering algorithms is affected by the number of clusters in the original one-way clustering from which they start. These results are shown in Table 2. The table shows the number of parts-of-speech produced by LP08 versus CDCC when started off with varying numbers of hard clusters

in the range 3 to 18. For each algorithm, the table shows (under *Any*) the number of distinct parts-of-speech (clusters) to which at least one word-frame occurrence was assigned during the categorization reported above, and also (under *1%*) the number of parts-of-speech such that at least one percent of the total number of word-frame combinations were assigned to that part-of-speech. The results under *Any* show that, as

	LP08		CDCC	
K	**Any**	**1%**	**Any**	**1%**
3	3	3	3	3
6	5	3	4	3
9	9	4	5	3
12	12	6	9	3
15	15	6	10	3
18	18	7	9	3

Table 2. Number of parts-of-speech used during categorization for three co-clustering algorithms, for varying *K* = number of clusters produced in initial one-way clustering. *Any* = number of parts-of-speech that account for at least one frame-word instance; *1%* = number of parts-of-speech that account for at least 1% of instances. *LP08* = replication of Leibbrandt & Powers (2008); *CDCC* = Conflict-Driven Co-Clustering.

the number of initial clusters grew, so too did the number of clusters that were used at least once during categorization, so that the algorithms were rather badly prone to proliferation of parts-of-speech when started with a large number of initial clusters, although CDCC was more conservative than LP08, and managed to discard many of the original clusters. However, the results for *1%* are more encouraging. Both algorithms, even when started with several candidate clusters in the one-way clustering, managed to eliminate the minor clusters to some extent, and redistribute their members into the larger parts-of-speech. It is particularly noteworthy that for CDCC, only three clusters were used for more than 1% of all instances. Inspection of the details of categorization showed that the CDCC algorithm managed to discover three clusters that seemed to correspond closely to the three major English parts-of-speech of Noun, Verb and Adjective. These categories appeared to be such a salient feature of the data for CDCC that they were able to 'self-organize' during runs of the algorithm from various one-way clustering starting points. This robust induction of the main English parts-of-speech is a striking advantage of CDCC over LP08.

It may be argued that the number of classes produced by the algorithm are too few to provide a basis for part-of-speech induction. To some extent this is a consequence of the seeding algorithm chosen. The frames used by Leibbrandt & Powers (2008, 2010) tended to support mostly open-class word fillers; nouns, verbs and adjectives made up respectively 52%, 25% and 10% of the total number of tokens that served as fillers in their frames, for a total of 87%. Arguably, this may be seen as desirable: for a child learning a language, knowledge of the open classes is more useful for learning novel words than knowledge of the closed classes. On the other hand, the lack of a category of adverbs may be regarded as a shortcoming of the original work by Leibbrandt & Powers. Nevertheless, the CDCC algorithm was able to robustly identify the main classes represented in the co-occurrence matrix.

4 Discussion

The CDCC algorithm has been shown to achieve similar categorization performance to some earlier models of part-of-speech induction. The most striking advantage has been that CDCC is able to "hone in" on the three main parts-of-speech. We suggest that this is due to the conservative nature of conflict resolution: by tallying the strength of evidence for a particular category in terms of the number of votes it receives, weaker categories are not able to cast sufficient numbers of votes to change word or frame allocations. Importantly, this means that in subsequent iterations, when conflicts are recalculated and votes cast once more, allocations of particular words or frames to these minor categories are more likely to be swamped by the additional allocations previously added to the major categories, so that the initially stronger categories become stronger as the algorithm executes while the weaker categories all but disappear. This is an important feature of the algorithm, because the original clustering step from which both CDCC and Leibbrandt & Powers (2008) begin is unconstrained in the number of clusters it produces; this is a parameter of the system, but it is a relatively unimportant one in the case of CDCC because the algorithm self-organizes around the major categories.

While the CDCC algorithm performs similarly to other established work while taking a radically different approach, several issues remain to be investigated. One of the potential strengths of CDCC is that it treats category membership in a discrete or symbolic way, rather than graded, as in Leibbrandt & Powers (2008). It remains to be seen whether such a treatment provides specific benefits in resolving ambiguity when dealing with words or frames that can belong to multiple categories.

CDCC has been formulated here as combining distributional information about the word type and the frame type in order to produce a part-of-speech allocation. However, the algorithm can be viewed more generally as a method to combine or fuse more than one source of information together, and hence can be applied to distributional, phonological, semantic or any other forms of linguistic information.

As it has been formulated here as a batch process, the CDCC algorithm can be regarded as addressing only the computational level of the problem of part-of-speech induction in language acquisition. Additional work would be required to attempt to address the algorithmic or implementational levels by turning the algorithm into a fully incremental learner (e.g., Parisien, Fazly & Stevenson, 2008; Chrupala & Alishahi, 2010). A simple variant of the CDCC algorithm could be one that simply processes the corpus in order, and in the case of a conflict between word and frame, stores the occurrence as evidence that the membership of either the word or frame should be altered, and in what way. When the accumulated evidence for a specific change of membership exceeds a threshold (e.g. when a certain number of votes have been cast to add membership of a particular cluster to a word or frame), the membership is added. It would remain to be determined empirically whether this iterative variant is still able to exhibit the same categorization performance and the property of robustness shown above for the batch CDCC algorithm.

References

Adriaans, P. (1992). *Language Learning from a Categorial Perspective.* Unpublished PhD thesis, University of Amsterdam.

Adriaans, P. (1999). *Learning Shallow Context-Free Languages under Simple Distributions* (Technical Report No. PP-1999-13): Institute for Logic, Language and Computation, University of Amsterdam.

Berg-Kirkpatrick, T., Côté, A.B., DeNero, J. & Klein, D. (2010). Painless unsupervised learning with features. *Proceedings of NAACL 2010*, 582–590.

Bybee, J. L. (1985). *Morphology: a study of the relation between meaning and form*: John Benjamins.

Bybee J. L. (2006). From usage to grammar: the mind's response to repetition. *Language*, *82*,711–733.

Christodoulopoulos, C., Goldwater, S. & Steedman, M. (2010). Two Decades of Unsupervised POS induction: How far have we come? *Proceedings of EMNLP 2010*, 575-584.

Chrupala, G. & Alishahi, A. (2010). Online Entropy-based Model of Lexical Category Acquisition. *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL-2010).*

Clark, A. (2000). Inducing syntactic categories by context distribution clustering. *Proceedings of the Conference on Natural Language Learning (CONLL-2000)*, 91–94.

Erkelens, M. (2008). Restrictions of frequent frames as cues to categories: the case of Dutch. *Supplement to the Proceedings of the 32nd Boston University Conference on Language Development (BUCLD 32).*

Freitag, D. (2004). Toward unsupervised whole-corpus tagging. *Proceedings of COLING-04*, 357-363.

Goldwater, S. & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of ACL 2007*, 744–751,

Leibbrandt, R. E., & Powers, D. M. W. (2008). Grammatical category induction using lexically-based templates. *Supplement to the Proceedings of the 32nd Boston University Conference on Language Development (BUCLD 32)*.

Leibbrandt, R.E. & Powers, D.M. (2010). Frequent Frames as Cues to Part-of-Speech in Dutch: Why Filler Frequency Matters. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2680-2685.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (3rd ed. Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum.

Macnamara, J. (1982). *Names for things: a study of child language*. Cambridge, MA: MIT Press.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics, 1*(1), 24-45.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's Language* (Vol. 2). New York: Gardner Press.

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

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Moon, T., Erk, K. & Baldridge, J. (2010) Crouching Dirichlet, Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation. *Proceedings of EMNLP 2010*, 196-206.

Parisien, C., Fazly, A. & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. *Proceedings of the 12th Conference on Computational Natural Language Learning, (CONLL-2008).*

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Powers, D. M. W. (2003). *Recall and precision versus the Bookmaker*. Paper presented at the 4th International Conference on Cognitive Science (ICCS).

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

Rosenberg, A. & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*, pp. 410-420.

Schütze, H. (1995) Distributional part-of-speech tagging. *Proceedings of EACL-95*.

Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge University Press.

St. Clair, M.C., Monaghan, P. & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition, 116*, 341-360.

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

Van Mechelen, I. & De Boeck, P. (2004). Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research, 13*, 363-394.