each of them with a POS tag from the tagset. We found that all of the clusters are labeled as NN, VB, or JJ. The reason is that the clustered words are mostly root words. We then merge all the clusters labeled with the same POS tag, yielding only three "big" clusters. Note that these "big" clusters are *soft* clusters, since a word can belong to more than one of them. For instance, "cool" can combine with "s" or "ing" to form a VB, and it can also combine with "er" or "est" to form a JJ.

Generating sub-clusters. Recall that each "big" cluster contains a set of suffixes and also a set of words that combines with those suffixes. Now, for each "big" cluster c, we create one sub-cluster c_x for each suffix x that appears in c. Then, for each word w in c, we use our unsupervised morphological analyzer to generate $w+x$ and add the surface form to the corresponding sub-cluster.

Labeling the sub-clusters. Finally, we manually label each sub-cluster with a POS tag from our tagset. For example, all the words ending in "ing" will be labeled as VBG. As before, we merge two clusters if they are labeled with the same POS tag. The resulting clusters are our morphologically formed clusters.

5 Purifying the Seed Set

The clusters formed thus far cannot be expected to be perfectly accurate, since (1) our unsupervised morphological analyzer is not perfect, and (2) morphology alone is not always sufficient for determining the POS of a word. In fact, we found that many adjectives are mislabeled as nouns for both languages. For instance, "historic" is labeled as a noun, since it combines with suffixes like "al" and "ally" that "accident" combines with. In addition, many words are labeled with the POS that does not correspond to their most common word sense. For instance, while words like "chair", "crowd" and "cycle" are more commonly used as nouns than verbs, they are labeled as verbs by our clustering algorithm. The reason is that suffixes that typically attach to verbs (e.g., "s", "ed", "ing") also attach to these words. Such labelings, though not incorrect, are undesirable, considering the fact that these words are to be used as seeds to bootstrap our morphologically formed clusters in a distributional manner. For instance, since "chair" and "crowd" are distributionally similar to nouns, their presence in the verb clusters can potentially contaminate the clusters with nouns during the bootstrapping process. Hence, for the purpose of effective bootstrapping, we also consider these words "mislabeled".

To identify the words that are potentially mislabeled, we rely on the following assumption: words that are morphologically similar should also be distributionally similar and vice versa. Based on this assumption, we propose a *purification* method that posits a word w as potentially mislabeled (and therefore should be removed or relabeled) if the POS of w as predicted using distributional information differs from that as determined by morphology.

The question, then, is how to predict the POS tag of a word using distributional information? Our idea is to use "supervised" learning, where we train and test on the seed set. Conceptually, we (1) train a *multi-class* classifier on the morphologically labeled words, each of which is represented by its context vector, and (2) apply the classifier to relabel the *same* set of words. If the new label of a word w differs from its original label, then morphology and context disagree upon the POS of w; and as mentioned above, our method then determines that the word is potentially misclassified. Note, however, that (1) the training instances are not perfectly labeled and (2) it does not make sense to train a classifier on data that is seriously mislabeled. Hence, we make the assumption that a large percentage (> 70%) of the training instances is correctly labeled[6], and that our method would work with a training set labeled at this level of accuracy. In addition, since we are training a classifier based on distributional features, we train and test on only *distributionally reliable* words, which we define to be words that appear at least five times in our corpus. Distributionally unreliable words will all be removed from the morphologically formed clusters, since we cannot predict their POS using distributional information.

In our implementation of this method, rather than train a multi-class classifier, we train a set of binary classifiers using SVM^{light} (Joachims, 1999) together with the distributional features for determining the POS tag of a given word.[7] More specifically, we train one classifier for each pair of

[6] An inspection of the morphologically formed clusters reveals that this assumption is satisfied for both languages.

[7] In this and all subsequent uses of SVM^{light}, we set all the training parameters to their default values.