emphasis on resource-scarce languages, our approach does not rely on any language resources. In particular, the morphological information that it exploits is provided by an unsupervised morphological analyzer.

It is perhaps not immediately clear why morphological information would play a crucial role in the induction process, especially since the distributional approach has achieved considerable success for English POS induction (see Lamb (1961), Schütze (1995) and Clark (2000)). To understand the role and significance of morphology, it is important to first understand why the distributional approach works well for English. Recall from the above that the distributional approach assumes that the information encoded in the context vector of each word, which typically consists of the 250 most frequent words of a given language, is sufficient for accurately clustering the words. This approach works well for English because the most frequent English words are composed primarily of closed-class words such as "to" and "is", which provide strong clues to the POS of the target word. However, this assumption is not necessarily valid for fairly free word order and highly inflectional languages such as Bengali. The reason is that (1) co-occurrence statistics collected from free word order languages are not as reliable as those from fixed word order languages; and (2) many of the closed-class words that appear in the context vector for English words are realized as inflections in Bengali. The absence of these highly informative words implies that the context vectors may no longer capture sufficient information for accurately clustering Bengali words, and hence the use of morphological information becomes particularly important for unsupervised POS induction for these inflectional languages.

We will focus primarily on labeling *open-class* words with their POS tags. Our decision is motivated by the fact that closed-class words generally comprise a small percentage of the lexical items of a language. In Bengali, the percentage of closed-class words is even smaller than that in English: as mentioned before, many closed-class words in English are realized as suffixes in Bengali.

Although our attempt to incorporate morphological information into the distributional POS induction framework was originally motivated by inflectional languages, experimental results show that our approach works well for both English and Bengali, suggesting its applicability to both morphologically impoverished languages and highly inflectional languages. Owing to the lack of publicly available resources for Bengali, we manually created a 5000-word Bengali lexicon for evaluation purposes. Hence, one contribution of our work lies in the creation of an annotated dataset for Bengali. By making this dataset publicly available[1], we hope to facilitate the comparison of different unsupervised POS induction algorithms and to stimulate interest in Bengali language processing.

The rest of the paper is organized as follows. Section 2 discusses related work on unsupervised POS induction. Section 3 describes our tagsets for English and Bengali. The next three sections describe the three steps of our bootstrapping approach: cluster the words using morphological information (Section 4), remove potentially mislabeled words from each cluster (Section 5), and bootstrap each cluster using a weakly supervised learner (Section 6). Finally, we present evaluation results in Section 7 and conclusions in Section 8.

2 Related Work

Several unsupervised POS induction algorithms have also attempted to incorporate morphological information into the distributional framework, but our work differs from these in two respects.

Computing morphological information. Previous POS induction algorithms have attempted to derive morphological information from dictionaries (Hajič, 2000) and knowledge-based morphological analyzers (Duh and Kirchhoff, 2006). However, these resources are generally not available for resource-scarce languages. Consequently, researchers have attempted to derive morphological information heuristically (e.g., Cucerzan and Yarowsky (2000), Clark (2003), Freitag (2004)). For instance, Cucerzan and Yarowsky (2000) posit a character sequence x as a suffix if there exists a sufficient number of distinct words w in the vocabulary such that the concatentations wx are also in the vocabulary. It is conceivable that such heuristically computed morphological information can be inaccurate, thus rendering the usefulness of a more accurate morphological analyzer. To address this problem, we exploit morphological information provided by an unsupervised word segmentation algorithm.

[1] See http://www.utdallas.edu/~sajib/posDatasets.html.