Tag	Description	Treebank tags
JJ	Adjective	JJ
JJR	Adjective, comparative	JJR
JJS	Adjective, superlative	JJS
NN	Singular noun	NN, NNP
NNS	Plural noun	NNS, NNPS
RB	Adverb	RB
VB	Verb, non-3rd ps. sing. present	VB, VBP
VBD	Verb, past tense or past participle	VBD, VBN
VBG	Verb, gerund/present participle	VBG
VBZ	Verb, 3rd ps. sing. present	VBZ

Table 1: The English tagset

Tag	Description	Examples
JJ	Adjective	vhalo, garam, kharap
NN	Singular noun	kanna, ridoy, shoshon
NN2	2nd order inflectional noun	dhopake, kalamtike
NN6	6th order inflectional noun	gharer, manusher
NN7	7th order inflectional noun	dhakai, barite, graame
NNP	Proper noun	arjun, ahmmad
NNS	Plural noun	manushgulo, pakhider
NNSH	Noun ending with "sh"	barish, jatrish
VB	Finite verb	kheyechi, krlam, krI
VBN	Non-finite verb	kre, giye, jete, kadte

Table 2: The Bengali tagset

Using morphological information. Perhaps due to the overly simplistic methods employed to compute morphological information, morphology has only been used as what Biemann (2006) called *add-on's* in existing POS induction algorithms, which remain primarily distributional in nature. In contrast, our approach more tightly integrates morphology into the distributional framework. As we will see, we train SVM classifiers using both morphological and distributional features to select seed words for our bootstrapping algorithm, effectively letting SVM combine these two sources of information and perform automatic feature weighting. Another appealing feature of our approach is that when labeling each unlabeled word with its POS tag, an SVM classifier also returns a numeric value that indicates how confident the word is labeled. This opens up the possibility of having a human improve our automatically constructed lexicon by manually checking those entries that are tagged with low confidence by an SVM classifier.

Recently, there have been attempts to perform (mostly) unsupervised POS tagging without relying on a POS lexicon. Haghighi and Klein's (2006) *prototype-driven* approach requires just a few prototype examples for each POS tag, exploiting these labeled words to constrain the labels of their distributionally similar words when training a generative log-linear model for POS tagging. Smith and Eisner (2005) train a log-linear model for POS tagging in an unsupervised manner using *contrastive estimation*, which seeks to move probability mass to a positive example *e* from its *neighbors* (i.e., negative examples created by perturbing *e*).

3 The English and Bengali Tagsets

Given our focus on automatically labeling open class words, our English and Bengali tagsets are designed to essentially cover all of the open-class words. Our English tagset, which is composed of ten tags, is shown in Table 1. As we can see, a tag in our tagset can be mapped to more than one Penn Treebank tags. For instance, we use the tag "NN" for both singular and plural common nouns. Our decision of which Penn Treebank tags to group together is based on that of Schütze (1995).

Our Bengali tagset, which also consists of ten tags, is adapted from the one proposed by Saha et al. (2004) (see Table 2). It is worth noting that unlike English, we assign different tags to Bengali proper nouns and common nouns. The reason is that for English, it is not particularly crucial to distinguish the two types of nouns during POS induction, since they can be distinguished fairly easily using heuristics such as initial capitalization. For Bengali, such simple heuristics do not exist, as the Bengali alphabet does not have any upper and lower case letters. Hence, it is important to distinguish Bengali proper nouns and common nouns during POS induction.

4 Clustering the Morphologically Similar Words

As mentioned before, our approach aims to more tightly integrate morphological information into the distributional POS induction framework. In fact, our POS induction algorithm begins by clustering the *morphologically similar* words (i.e., words that combine with the same set of suffixes). The motivation for clustering morphologically similar words can be attributed to our hypothesis that words having similar POS should combine with a similar set of suffixes. For instance, verbs in English combine with suffixes like "ing", "ed" and "s", whereas adjectives combine with suffixes like "er" and "est". Note, however, that the suffix "s" can attach to both verbs and nouns in English, and so it is not likely to be a useful feature for identify-