# Crouching Dirichlet, Hidden Markov Model: Unsupervised POS Tagging with Context Local Tag Generation

**Taesun Moon, Katrin Erk, and Jason Baldridge**
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
{tsmoon,katrin.erk,jbaldrid}@mail.utexas.edu

## Abstract

We define the crouching Dirichlet, hidden Markov model (CDHMM), an HMM for part-of-speech tagging which draws state prior distributions for each local document context. This simple modification of the HMM takes advantage of the dichotomy in natural language between content and function words. In contrast, a standard HMM draws all prior distributions once over all states and it is known to perform poorly in unsupervised and semi-supervised POS tagging. This modification significantly improves unsupervised POS tagging performance across several measures on five data sets for four languages. We also show that simply using different hyperparameter values for content and function word states in a standard HMM (which we call HMM+) is surprisingly effective.

## 1 Introduction

Hidden Markov Models (HMMs) are simple, versatile, and widely-used generative sequence models. They have been applied to part-of-speech (POS) tagging in supervised (Brants, 2000), semi-supervised (Goldwater and Griffiths, 2007; Ravi and Knight, 2009) and unsupervised (Johnson, 2007) training scenarios. Though discriminative models achieve better performance in both semi-supervised (Smith and Eisner, 2005) and supervised (Toutanova et al., 2003) learning, there has been only limited work on unsupervised discriminative sequence models (e.g., on synthetic data and protein sequences (Xu et al., 2006)), and none to POS tagging.

The tagging accuracy of purely unsupervised HMMs is far below that of supervised and semi-supervised HMMs; this is unsurprising as it is still not well understood what kind of structure is being found by an unconstrained HMM (Headden III et al., 2008). However, HMMs are fairly simple directed graphical models, and it is straightforward to extend them to define alternative generative processes. This also applies to linguistically motivated HMMs for recovering states and sequences that correspond more closely to those implicitly defined by linguists when they label sentences with parts-of-speech.

One way in which a basic HMM's structure is a poor model for POS tagging is that there is no inherent distinction between (open-class) content words and (closed-class) function words. Here, we propose two extensions to the HMM. The first, HMM+, is a very simple modification where two different hyperparameters are posited for content states and function states, respectively. The other is the *crouching Dirichlet, hidden Markov model* (CDHMM), an extended HMM that captures this dichotomy based on the statistical evidence that comes from context. Content states display greater variance across local context (e.g. sentences, paragraphs, documents), and we capture this variance by adding a component to the model for content states that is based on latent Dirichlet allocation (Blei et al., 2003). This extension is in some ways similar to the LDAHMM of Griffiths et al. (2005). Both models are composite in that two distributions do not mix with each other. Unlike the LDAHMM, the generation of content states is folded into the CDHMM process.

We compare the HMM+ and CDHMM against a basic HMM and LDAHMM on POS tagging on a more extensive and diverse set of languages than previous work in monolingual unsupervised POS tagging: four languages from three families (*Germanic*: English and German; *Romance*: Portuguese;



*Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 196–206,
MIT, Massachusetts, USA, 9-11 October 2010. 

and *Mayan*: Uspanteko). The CDHMM easily outperforms all other models, including HMM+, across three measures (accuracy, F-score, and variation of information) for unsupervised POS tagging on most data sets. However, the HMM+ is surprisingly competitive, outperforming the basic HMM and LDAHMM, and rivaling or even passing the CDHMM on some measures and data sets.

## 2 Background

The Bayesian formulation for a basic HMM (Goldwater and Griffiths, 2007) is:

$$\begin{aligned} \psi_t|\xi &\sim \text{Dir}(\xi) \\ \delta_t|\gamma &\sim \text{Dir}(\gamma) \\ w_i|t_i = t &\sim \text{Mult}(\psi_t) \\ t_i|t_{i-1} = t &\sim \text{Mult}(\delta_t) \end{aligned}$$

Dir is the conjugate Dirichlet prior to Mult (a multinomial distribution). The state transitions are generated by $\text{Mult}(\delta_t)$ whose prior $\delta_t$ is generated by $\text{Dir}(\gamma)$ with a symmetric (i.e. uniform) hyperparameter $\gamma$. Emissions are generated by $\text{Mult}(\psi_t)$ with a prior $\psi_t$ generated by $\text{Dir}(\xi)$ with a symmetric hyperparameter $\xi$. Hyperparameter values smaller than one encourage posteriors that are peaked, with smaller values increasing this concentration. It is not necessary that the hyperparameters be symmetric, but this is a common approach when one wants to be naïve about the data. This is particularly appropriate in unsupervised POS tagging with regard to novel data since there won't be *a priori* grounds for favoring certain distributions over others.

There is considerable work on extensions to HMM-based unsupervised POS tagging (see §6), but here we concentrate on the LDAHMM (Griffiths et al., 2005), which models topics and state sequences jointly. The model is a composite of a probabilistic topic model and an HMM in which a single state is allocated for words generated from the topic model. A strength of this model is that it is able to use less supervision than previous topic models since it does not require a stopword list. While the topic model component still uses the bags-of-words assumption, the joint model infers which words are more likely to carry topical content and which words are more likely to contribute to the local sequence. This model is competitive with a standard topic model, and its output is also competitive when compared with a standard HMM. However, Griffiths et al. (2005) note that the topic model component inevitably loses some finer distinctions with respect to parts-of-speech. Though many content states such as adjectives, verbs, and nouns can vary a great deal across documents, the topic state groups these words together. This leads to assignment of word tokens to clusters that are a poorer fit for POS tagging. This paper shows that a model that conflates the LDAHMM topics with content states can significantly improve POS tagging.

## 3 Models

We aim to model the fact that in many languages words can generally be grouped into function words and content words and that these groups often have significantly different distributions. There are few function words and they appear frequently, while there are many content words appearing infrequently. Another difference in distribution is often implied in information retrieval by the use of stopword filters and *tf-idf* values to remove or reduce the influence of words which occur frequently but have low variance (i.e. their global probability is similar to their local probability in a document).

A difference in distribution is also revealed when the parts-of-speech are known. When no smoothing parameters are added, the joint probability of a word that is not 'the' or 'a' occurring with a `DT` tag (in the Penn Treebank) is almost always zero. Similarly peaked distributions are observed for other function categories such as `MD` and `CC`. On the other hand, the joint probability of any word occurring with `NN` is much less likely to be zero and the distribution is much less likely to be peaked.

We attempt to account for these two distributional properties—that certain words have higher variance across contexts (e.g. a document) and that certain tags have more peaked emission distributions—in a sequence model. To do this, we define the *crouching Dirichlet, hidden Markov model*[1] (CDHMM). This model, like LDAHMM, captures items of high variance across contexts, but it does so without losing

[1]We call our model a "crouching Dirichlet" model since it involves a Dirichlet prior that generates distributions for certain states as if it were "crouching" on the side.

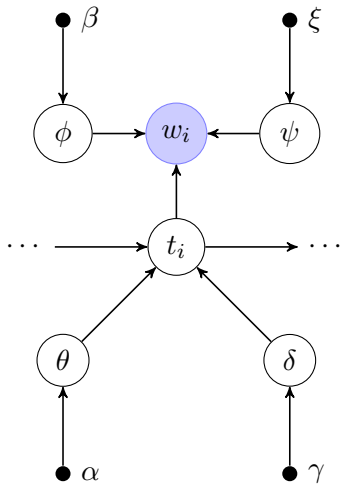


Figure 1: Graphical representation of relevant variables and dependencies at a given time step $i$. Observed word $w_i$ is dependent on hidden state $t_i$. Edges to priors $\theta, \phi, \psi$ may or may not be activated depending on the value of $t_i$. The edge to transition prior $\delta$ is always activated. Hyperparameters to priors are represented by dots. See §3.1 for details.

sequence distinctions, namely, a given word's local function via its part-of-speech. We also define the HMM+, a simple adaptation of a basic HMM which accounts for the latter property by using different priors for emissions from content and function states.

### 3.1 CDHMM

The CDHMM incorporates an LDA-like module to its graphical structure in order to capture words and tags which have high variance across contexts. Such tags correspond to content states. Like the LDAHMM, the model is composite in that distributions over a single random variable are composed of several different distribution functions which depend on the value of the underlying variable.

We posit the following model (see fig. 1 for a diagram of dependencies and all variables involved at a single time step). We observe a sequence of tokens $\mathbf{w}=(w_1, \ldots, w_N)$ that we assume is generated by an underlying state sequence $\mathbf{t}=(t_1, \ldots, t_N)$ over a state alphabet $T$ with first order Markov dependencies. $T$ is a union of disjoint content states $C$ and function states $F$. In this composite model, the priors for the emission and transition for each step in the sequence depend on whether state $t$ at step $i$ is $t \in C$ or $t \in F$. If $t \in C$, the word emission is dependent on $\phi$ (the content word prior) and the state transition is dependent on $\theta$ (the "topic" prior) and $\delta$ (the transition prior). If $t \in F$, the word emission probability is dependent on $\psi$ (the function word prior) and the state transition on $\delta$ (again, the transition prior). Therefore, if $t \in F$, the transition and emission structure is identical to the standard Bayesian HMM.

To elaborate, three prior distributions are defined globally for this model: (1) $\delta_t$, the transition prior such that $p(\hat{t}|t, \delta_t) = \delta_{\hat{t}|t}$ (2) $\psi_t$, the function word prior such that $p(w|t, \psi_t) = \psi_{w|t}$ (3) $\phi_t$, the content word prior such that $p(w|t, \phi_t) = \phi_{w|t}$. Locally for each context $d$ (documents in our case), we define $\theta_d$, the topic prior such that $p(t|\theta_d) = \theta_{t|d}$ for $t \in C$.

The generative story is as follows:

1. For each state $t \in T$
   (a) Draw a distribution over states $\delta_t \sim \text{Dir}(\gamma)$
   (b) If $t \in C$, draw a distribution over words $\phi_t \sim \text{Dir}(\beta)$
   (c) If $t \in F$, draw a distribution over words $\psi_t \sim \text{Dir}(\xi)$
2. For each context $d$
   (a) Draw a distribution $\theta_d \sim \text{Dir}(\alpha)$ over states $t \in C$
   (b) For each word $w_i$ in $d$
      i. draw $t_i$ from $\delta_{t_{i-1}} \circ \theta_d$
      ii. if $t_i \in C$, then draw $w_i$ from $\phi_{t_i}$, else draw $w_i$ from $\psi_{t_i}$

For each context $d$, we draw a prior distribution $\theta_d$—formally identical to the LDA topic prior—that is defined only for the states $t \in C$. This prior is then used to weight the draws for states at each word, from $\delta_{t_{i-1}} \circ \theta_d$, where we have defined the vector valued operation $\circ$ as follows:

$$(\delta_{t_{i-1}} \circ \theta_d)_{t_i} = \begin{cases} \frac{1}{Z}\delta_{t_i|t_{i-1}} \cdot \theta_{t_i|d} & t_i \in C \\ \frac{1}{Z}\delta_{t_i|t_{i-1}} & t_i \in F \end{cases}$$

where $(\delta_{t_{i-1}} \circ \theta_d)_{t_i}$ is the element corresponding to state $t_i$ in the vector $\delta_{t_{i-1}} \circ \theta_d$. $Z$ is a normalization constant such that the probability mass sums to one.

$$p(t_i|\mathbf{t}_{-i},\mathbf{w}) \propto \begin{cases} \frac{N_{w_i|t_i}+\beta}{N_{t_i}+W\beta}\ \frac{N_{t_i|d_i}+\alpha}{N_{d_i}+C\alpha}\ \frac{\left(N_{t_i|t_{i-1}}+\gamma\right)\left(N_{t_{i+1}|t_i}+\mathrm{I}[t_{i-1}=t_i=t_{i+1}]+\gamma\right)}{N_{t_i}+T\gamma+\mathrm{I}[t_i=t_{i-1}]} & t_i \in C \\ \frac{N_{w_i|t_i}+\xi}{N_{t_i}+W\xi}\ \frac{\left(N_{t_i|t_{i-1}}+\gamma\right)\left(N_{t_{i+1}|t_i}+\mathrm{I}[t_{i-1}=t_i=t_{i+1}]+\gamma\right)}{N_{t_i}+T\gamma+\mathrm{I}[t_i=t_{i-1}]} & t_i \in F \end{cases}$$

Figure 2: Conditional distribution for $t_i$ in the CDHMM.

The important thing to note is that the draw for states at each word is proportional to a *composite* of (a) the product of the individual elements of the topic and transition priors when $t_i \in C$ and (b) the transition priors when $t_i \in F$. The draw is proportional to the product of topic and transition priors when $t_i \in C$ because we have made a product of experts (PoE) factorization assumption (Hinton, 2002) for tractability and to reduce the size of our model. Without such an assumption, the transition parameters would lie in a partitioned space of size $O(|C|^4)$ as opposed to $O(|T|^2)$ for the current model. Furthermore, this combination of a composite hidden state space with a product of experts assumption allows us to capture high variance for certain states.

To summarize, the CDHMM is a composite model where both the observed token and the hidden state variable are composite distributions. For the hidden state, this means that there is a "topical" element with high variance across contexts that is embedded in the state sequence for a subset of events. We embed this element through a PoE assumption where transitions into content states are modeled as a product of the transition probability and the local probability of the content state.

**Inference.** We use a Gibbs sampler (Gao and Johnson, 2008) to learn the parameters of this and all other models under consideration. In this inference regime, two distributions are of particular interest. One is the posterior density and the other is the conditional distribution, neither of which can be learned in closed form.

Letting $\Lambda = (\theta, \delta, \phi, \psi)$ and $h = (\alpha, \beta, \gamma, \xi)$, the posterior density is given as

$$p(\Lambda|\mathbf{w},\mathbf{t};h) \propto p(\mathbf{w},\mathbf{t}|\Lambda)p(\Lambda;h)$$

Note that $p(\mathbf{w},\mathbf{t}|\Lambda)$ is equal to

$$\prod_d^D \prod_i^{N_d} \left(\phi_{w_i|t_i}\theta_{t_i|d}\delta_{t_i|t_{i-1}}\right)^{\mathrm{I}[t_i\in C]} \left(\psi_{w_i|t_i}\delta_{t_i|t_{i-1}}\right)^{\mathrm{I}[t_i\in F]} \quad (1)$$

where $\mathrm{I}[\cdot]$ is the indicator function, $D$ is the number of documents in the corpus and $N_d$ is the number of tokens in document $d$.

Another important measure is the conditional distribution which is conditioned on all the random variables except the hidden state variable of interest and which is derived by integrating out the priors:

$$p(t_i|\mathbf{t}_{-i},\mathbf{w};h) \propto p(t_i|\mathbf{t}_{-i};h)p(w_i|\mathbf{t},\mathbf{w}_{-i};h) \quad (2)$$

where $\mathbf{t}_{-i}$ is the joint random variable $\mathbf{t}$ without $t_i$ and $\mathbf{w}_{-i}$ is $\mathbf{w}$ without $w_i$.

There are two well-known approaches to conducting Gibbs sampling for HMMs. The default method is to sample $\Lambda$ based on the posterior, then sample each $t_i$ based on the conditional distribution. Another approach is to sample directly from the conditional distribution without sampling from the posterior since the conditional distribution incorporates the posterior through integration. This is called a collapsed Gibbs sampler, which is the method employed for the models in this study.

The full conditional distribution for tag transitions for the Gibbs sampler is given in Figure 2. At each time step, we decrement all counts for the current value of $t_i$, sample a new value for $t_i$ from a multinomial proportional to the conditional distribution and assign that value to $t_i$. $\beta, \xi$ are the hyperparameters for the word emission priors of the content states and function states, respectively. $\gamma$ is the hyperparameter for the state transition priors. $\alpha$ is the hyperparameter for the state prior given that it is in some context $d$. Note that we have overridden notation so

that $C$ and $T$ here refer to the size of the alphabet. $W$ is the size of the vocabulary. Notation such as $N_{t_i|t_{i-1}}$ refers to the counts of the events indicated by the subscript, minus the current token and tag under consideration. $N_{t_i|t_{i-1}}$ is the number of times $t_i$ has occurred after $t_{i-1}$ minus the tag for $w_i$. $N_{w_i|t_i}$ is the number of times $w_i$ has occurred with $t_i$ minus the current value. $N_{t_i}$ and $N_{d_i}$ are the counts for the given tag and document minus the current value.

In its broad outline, the CDHMM is not much more complicated than an HMM since the decomposition (eqn. 1) is nearly identical to that of an HMM with the exception that conditional probabilities for a subset of the states—the content states—are local. An inference algorithm can be derived that involves no more than adding a single term to the standard MCMC algorithm for HMMs (see Figure 2).

### 3.2 HMM+

The CDHMM explicitly posits two different types of states: function states and content states. Having made this distinction, there is a very simple way to capture the difference in emission distributions for function and content states within an otherwise standard HMM: posit different hyperparameters for the two types. One type has a small hyperparameter to model a sparse distribution for function words and the other has a relatively large hyperparameter to model a distribution with broader support. This extension, which we refer to as HMM+, provides an important benchmark to compare with the CDHMM to see how much is gained by its additional ability to model the fact that function words occur frequently but have low variance across contexts.

As with the CDHMM, we use Gibbs sampling to estimate the model parameters while holding the two different hyperparameters fixed. The conditional distribution for tag transitions for this model is identical to that in fig. 2 except that it does not have the second term $\frac{N_{t_i|d_i}+\alpha}{N_{d_i}+C\alpha}$ in the first case where $t_i \in C$.

We are not aware of a published instance of such an extension to the HMM—which our results show to be surprisingly effective. Goldwater and Griffiths (2007) posits different hyperparameters for individual states, but not for different groups of states.

| corpus | tokens | docs | avg. | tags |
|---|---|---|---|---|
| WSJ | 974254 | 1801 | 541 | 43 |
| Brown | 797328 | 343 | 2325 | 80 |
| Tiger | 447079 | 1090 | 410 | 58 |
| Floresta | 197422 | 1956 | 101 | 19 |
| Uspanteko | 70125 | 29 | 2418 | 83 |

Table 2: Number of tokens, documents, average tokens per document and total tag types for each corpus.

## 4 Data and Experiments

**Data.** We use five datasets from four languages (English, German, Portuguese, Uspanteko) for evaluating POS tagging performance.

- *English*: the Brown corpus (Francis et al., 1982) and the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1994).
- *German*: the Tiger corpus (Brants et al., 2002).
- *Portuguese*: the full Bosque subset of the Floresta corpus (Afonso et al., 2002).
- *Uspanteko* (an endangered Mayan language of Guatemala): morpheme-segmented and POS-tagged texts collected and annotated by the OKMA language documentation project (Pixabaj et al., 2007); we use the cleaned-up version described in Palmer et al. (2009).

Table 2 provides the statistics for these corpora.

We lowercase all words, do not remove any punctuation or *hapax legomena*, and we do not replace numerals with a single identifier. Due to the nature of the models, document boundaries are retained.

**Evaluation** We report values for three evaluation metrics on all five corpora, using their full tagsets.

- *Accuracy*: We use a greedy search algorithm to map each unsupervised tag to a gold label such that accuracy is maximized. We evaluate on a **1-to-1** mapping between unsupervised tags and gold labels, as well as many-to-1 (**M-to-1**), corresponding to the evaluation mappings used in Johnson (2007). The 1-to-1 mapping provides a stricter evaluation. The many-to-one mapping, on the other hand, may be more adequate as unsupervised tags tend to be more fine-grained than

| | Model | Accuracy | | Pairwise P/R Scores | | | VI |
|---|---|---|---|---|---|---|---|
| | | 1-to-1 | M-to-1 | P | R | F | |
| WSJ (50) | HMM | 0.34 (0.01) | 0.49 (0.03) | **0.51** (0.03) | 0.19 (0.01) | 0.28 (0.01) | 3.72 (0.08) |
| | LDAHMM | 0.30 (0.04) | 0.45 (0.04) | 0.25 (0.07) | 0.27 (0.03) | 0.26 (0.04) | 3.64 (0.14) |
| | HMM+ | ***0.42*** (0.04) | 0.46 (0.05) | 0.24 (0.03) | **0.49** (0.03) | 0.32 (0.03) | ***2.65*** (0.15) |
| | CDHMM | ***0.44*** (0.01) | **0.58** (0.02) | 0.31 (0.01) | 0.43 (0.03) | **0.36** (0.02) | ***2.73*** (0.08) |
| Brown (50) | HMM | 0.32 (0.01) | 0.50 (0.02) | **0.60** (0.02) | 0.18 (0.00) | 0.28 (0.01) | 3.82 (0.05) |
| | LDAHMM | 0.28 (0.06) | 0.41 (0.08) | 0.25 (0.10) | 0.28 (0.05) | 0.25 (0.05) | 3.71 (0.21) |
| | HMM+ | 0.43 (0.06) | 0.48 (0.07) | 0.29 (0.05) | 0.50 (0.04) | ***0.37*** (0.05) | 2.63 (0.19) |
| | CDHMM | **0.48** (0.02) | **0.62** (0.02) | 0.32 (0.03) | **0.54** (0.04) | ***0.40*** (0.03) | **2.48** (0.06) |
| Tiger (50) | HMM | 0.29 (0.02) | 0.49 (0.02) | **0.49** (0.04) | 0.14 (0.01) | 0.22 (0.02) | 3.91 (0.06) |
| | LDAHMM | 0.31 (0.04) | 0.50 (0.04) | 0.26 (0.07) | 0.24 (0.02) | 0.25 (0.04) | 3.51 (0.11) |
| | HMM+ | 0.41 (0.08) | 0.44 (0.05) | 0.25 (0.05) | ***0.58*** (0.10) | 0.35 (0.06) | ***2.70*** (0.25) |
| | CDHMM | **0.47** (0.01) | **0.61** (0.02) | 0.45 (0.01) | ***0.58*** (0.03) | **0.50** (0.02) | ***2.72*** (0.04) |
| Usp. (50) | HMM | 0.36 (0.01) | ***0.49*** (0.02) | **0.39** (0.01) | 0.18 (0.00) | **0.25** (0.00) | 3.63 (0.04) |
| | LDAHMM | 0.35 (0.02) | 0.47 (0.02) | 0.26 (0.04) | 0.23 (0.03) | 0.24 (0.02) | 3.52 (0.09) |
| | HMM+ | 0.32 (0.02) | 0.35 (0.03) | 0.12 (0.02) | **0.52** (0.05) | 0.20 (0.02) | 3.13 (0.06) |
| | CDHMM | **0.39** (0.02) | ***0.50*** (0.02) | 0.16 (0.02) | 0.39 (0.03) | 0.23 (0.02) | **3.00** (0.06) |
| Flor. (50) | HMM | 0.30 (0.01) | 0.58 (0.03) | **0.62** (0.05) | 0.18 (0.01) | 0.28 (0.01) | 3.51 (0.06) |
| | LDAHMM | ***0.36*** (0.06) | 0.59 (0.04) | 0.55 (0.10) | 0.29 (0.07) | ***0.38*** (0.08) | 3.22 (0.15) |
| | HMM+ | 0.35 (0.04) | 0.52 (0.02) | 0.28 (0.04) | **0.43** (0.06) | ***0.34*** (0.04) | **2.58** (0.07) |
| | CDHMM | ***0.36*** (0.01) | **0.64** (0.02) | 0.37 (0.02) | 0.27 (0.01) | 0.31 (0.01) | 2.73 (0.05) |

Table 1: Evaluation on WSJ, Brown, Tiger, Floresta and Uspanteko for models with 50 states. For VI, lower is better

gold part-of-speech tags. In particular, they tend to form semantically coherent sub-classes of gold parts of speech.

- *Pairwise Precision and Recall*: Viewing tagging as a clustering task over tokens, we evaluate pairwise precision ($P$) and recall ($R$) between the model tag sequence ($M$) and gold tag sequence ($G$) by counting the true positives ($tp$), false positives ($fp$) and false negatives ($fn$) between the two and setting $P = tp/(tp + fp)$ and $R = tp/(tp + fn)$. $tp$ is the number of token pairs that share a tag in $M$ as well as in $G$, $fp$ is the number token pairs that share the same tag in $M$ but have different tags in $G$, and $fn$ is the number token pairs assigned a different tag in $M$ but the same in $G$ (Meila, 2007). We also provide the $f$-score which is the harmonic mean of $P$ and $R$.
- *Variation of Information (VI)*: The variation of information is an information theoretic metric that measures the amount of information lost and gained in going from tag sequence $M$ to $G$ (Meila, 2007). It is defined as $VI(M, G) = H(M) + H(G) - 2I(M, G)$ where $H$ denotes entropy and $I$ mutual information. Goldwater and Griffiths (2007) noted that this measure can point out models that have more consistent errors in the form of lower VI, even when accuracy figures are the same.

We also report learning curves on **M-to-1** with geometrically increasing training set sizes of 8, 16, 32, 64, 128, 256, 512, 1024, and all documents, or as many as possible given the corpus.

## 5 Experiments

In this section we discuss our parameter settings and experimental results.

### 5.1 Models and Parameters

We compare four different models:

- HMM: a standard HMM
- HMM+: an HMM in which the hyperparameters for the word emissions are asymmetric, such that content states have different word emission priors compared to function states.
- LDAHMM: an HMM with a distinguished state that generates words from a topic model (Griffiths et al., 2005)

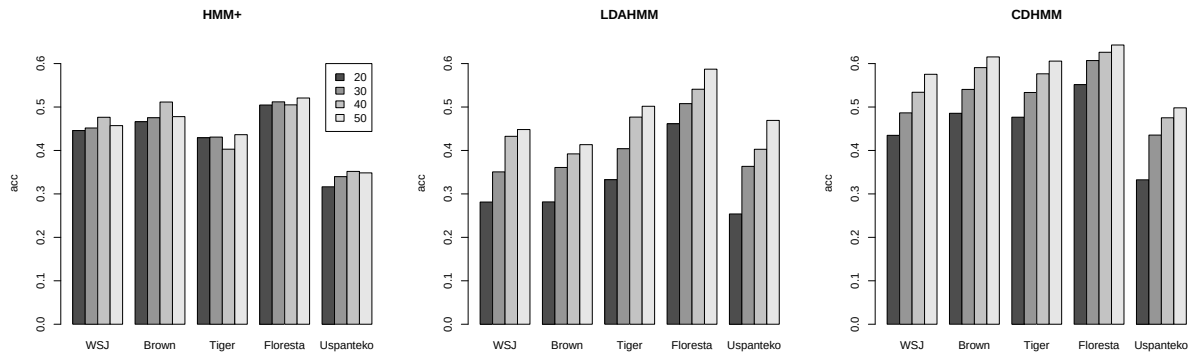


Figure 3: Averaged many-to-one accuracy on the full tagset for the models HMM+, LDAHMM, CDHMM when the number of states is set at 20, 30, 40 and 50 states.

- CDHMM: our HMM with context-based emissions, where the context used is the document

We implemented all of these models, ensuring performance differences are due to the models themselves rather than implementation details.

For all models, the transition hyperparameters $\gamma$ are set to $0.1$. For the LDAHMM and HMM all emission hyperparameters are set to 0.0001. These figures are the MCMC settings that provided the best results in Johnson (2007). For the models that distinguish content and function states (HMM+, CDHMM), we fixed the number of content states at 5 and set the function state emission hyperparameters $\xi = 0.0001$ and the content state emission hyperparameters $\beta = 0.1$. For the models with an LDA or LDA-like component (LDAHMM, CDHMM), we set the topic or content-state hyperparameter $\alpha = 1$.

For decoding, we use maximum posterior decoding to obtain a single sample after the required burn-in, as has been done in other unsupervised HMM experiments. We use this sample for evaluation.

### 5.2 Results

Results for all models on the full tagset are provided in table 1.[2] Each number is the mean accuracy of ten randomly initialized samples after a single chain burn-in of 1000 iterations. The model with a statistically significant ($p < 0.05$) best score for each measure and data set is given in plain bold. In cases where the differences for the best models are not significantly different from each other, but are significantly better from the others, the top model scores are given in bold italic.

CDHMM is extremely strong on the accuracy metric: it wins or ties for all datasets for both 1-to-1 and M-to-1 measures. For pairwise $f$-score, it obtains the best score for two datasets (WSJ and Tiger), and ties with HMM+ on Brown (we return to Uspanteko and Floresta below in an experiment that varies the number of states). For VI, HMM+ and CDHMM both easily outperform the other models, with CDHMM winning Brown and Uspanteko and HMM+ winning Floresta.

In the case of Uspanteko, the absolute difference in mean performance between models is smaller overall but still significant. This is due to the reduced variance between samples for all models. This is striking because the non-CDHMM models have much higher standard deviation on other corpora but have sharply reduced standard deviation only for Uspanteko. The most likely explanation is that the Uspanteko corpus is much smaller than the other corpora.[3] Nonetheless, CDHMM comes out strongest on most measures.

A simple baseline for accuracy is to choose the most frequent tag for all tokens; this gives accuracies of 0.14 (WSJ), 0.14 (Brown), 0.21 (Tiger), 0.20

[2] Similar results are obtained with reduced tagsets, as is commonly done in other work on unsupervised POS-tagging.

[3] which is interesting in itself since the weak law of large numbers implies that sample standard deviation decreases with sample size, which in our case is the number of tokens rather than the 10 samples under discussion

| | Model | Accuracy | | P/R Scores | | | VI |
|---|---|---|---|---|---|---|---|
| | | 1-to-1 | M-to-1 | P | R | F | |
| Usp. (100) | HMM | 0.36 (0.01) | 0.58 (0.01) | **0.56** (0.02) | 0.16 (0.00) | 0.25 (0.01) | 3.53 (0.04) |
| | LDAHMM | 0.35 (0.01) | 0.58 (0.02) | 0.45 (0.04) | 0.17 (0.01) | 0.24 (0.01) | 3.46 (0.06) |
| | HMM+ | 0.35 (0.02) | 0.41 (0.02) | 0.18 (0.01) | **0.36** (0.03) | 0.24 (0.01) | 3.25 (0.08) |
| | CDHMM | **0.40** (0.01) | **0.59** (0.01) | 0.25 (0.02) | 0.27 (0.02) | **0.26** (0.01) | **3.05** (0.03) |
| Flor. (20) | HMM | 0.31 (0.02) | 0.48 (0.03) | **0.40** (0.03) | 0.21 (0.01) | 0.28 (0.02) | 3.54 (0.10) |
| | LDAHMM | 0.35 (0.06) | 0.46 (0.06) | 0.27 (0.07) | 0.45 (0.08) | 0.33 (0.05) | 3.10 (0.10) |
| | HMM+ | 0.37 (0.04) | 0.50 (0.03) | 0.30 (0.02) | 0.45 (0.06) | 0.36 (0.03) | 2.62 (0.06) |
| | CDHMM | **0.44** (0.02) | **0.55** (0.02) | 0.30 (0.01) | **0.53** (0.03) | **0.39** (0.02) | **2.39** (0.07) |

Table 3: Evaluation for Uspanteko and Floresta. Experiments in this table use state sizes that correspond more closely to the size of the tag sets in the respective corpora.

(Floresta), and 0.11 (Uspanteko). Clearly, all of the models easily outperform this baseline.

**Number of states.** Figure 3 shows the change in accuracy for the different models for different corpora when the overall number of states is varied between 20 and 50. The figure shows results for **M-to-1**. All models with the exception of HMM+ show improvements as the number of states is increased. This brings up the valid concern (Clark, 2003; Johnson, 2007) that a model could posit a very large number of states and obtain high M-to-1 scores. However, it is neither the case here nor in any of the studies we cite. Furthermore, as is strongly suggested with HMM+, it does not seem as if all models will benefit from assuming a large number of states.

Looking at the results by number of states on VI and $f$-score for CDHMM(Figure 5), it is clear that Floresta displays the reverse pattern of all other data sets where performance monotonically deteriorates as state sizes are increased. Though the exact reason is unknown, we believe it is partially due to the fact that Floresta has 19 tags. We therefore wondered whether positing a state size that more closely approximated the size of the gold tag set performs better. Since the discrepancy is greatest for Uspanteko and Floresta, we present tabulated results for experiments with state settings of 100 and 20 states respectively (table 3). With the exception of VI (where lower is better) for Uspanteko, the scores generally improve when the model state size is closer to the gold size. **M-to-1** goes down for Floresta when 20 states are posited, but this is to be expected since this score is defined, to a certain extent, to do better with larger models.

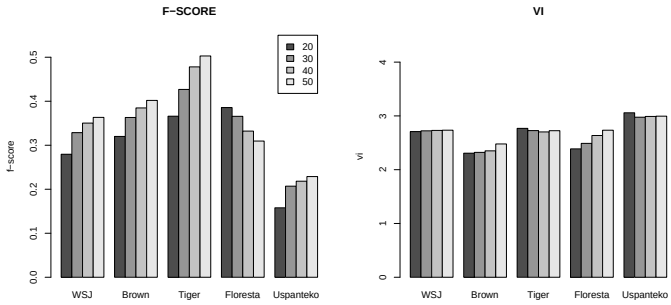


Figure 5: $f$-score and VI for CDHMM by number of states

**Variance.** As we average performance figures over ten runs for each model, it is also instructive to consider standard deviation across runs. Standard deviation is lowest for the CDHMM models and the vanilla HMM. Standard deviation is high for HMM+ and LDAHMM. This is not surprising for LDAHMM, since it has fifty topic parameters in addition to the number of states posited, and random initial conditions would have greater effect on the outcome than for the other models. It is unexpected, however, that HMM+ has high variance over different chains. The model shares the large content emission hyperparameter $\beta = 0.1$ with CDHMM. At this point, it can only be assumed that the additional LDA component acts as a regularization factor for CDHMM and reduced the volatility in having a large emission hyperparameter.

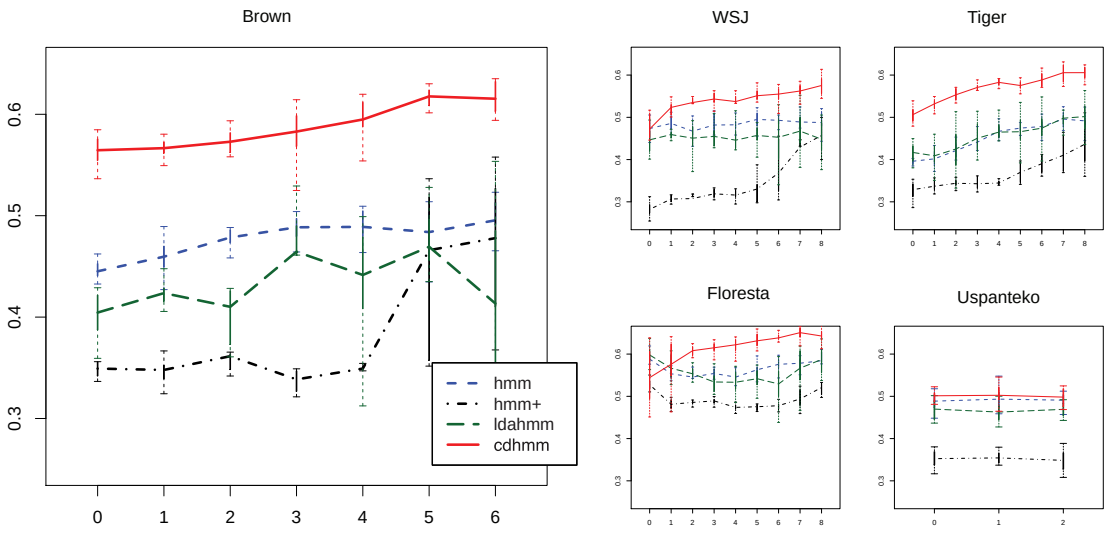


Figure 4: Learning curves on M-to-1 evaluation. The staples at each point represent two standard deviations.

**Learning curves** We present learning curves on different sizes of subcorpora in Figure 4. The graphs are box plots of the full M-1 accuracy figures on 10 randomly initialized training runs for seven subcorpora in Brown, nine in WSJ, Tiger, Floresta and three in Uspanteko.

Comparing the graphs, the performance of HMM+ shows the strongest improvement for English and German data as the amount of training data increases. Also, it is evident that CDHMM posts consistent performance gains across data sets as it trains on more data. This stands in opposition to HMM and LDAHMM which do not seem able to take advantage of more information for WSJ and Floresta. This suggests that performance for CDHMM and HMM+ could improve if the training corpora were augmented with out-of-corpus raw data. One exception to the consistent improvement over increased data is the performance of the models on Uspanteko, which uniformly flatline. One reason might be that the tags are labeled over segmented morphemes instead of words like the other corpora. Another could be that Uspanteko has a relatively large number of tags in a very small corpus.

## 6 Related work

Unsupervised POS tagging is an active area of research. Most recent work has involved HMMs. Given that an unconstrained HMM is not well understood in POS tagging, much work has been done on examining the mechanism and the properties of the HMM as applied to natural language data (Johnson, 2007; Gao and Johnson, 2008; Headden III et al., 2008). Conversely, there has also been work focused on improving the HMM as an inference procedure that looked at POS tagging as an example (Graca et al., 2009; Liang and Klein, 2009). Nonparametric HMMs for unsupervised POS tag induction (Snyder et al., 2008; Van Gael et al., 2009) have seen particular activity due to the fact that model size assumptions are unnecessary and it lets the data "speak for itself."

There is also work on alternative unsupervised models that are not HMMs (Schütze, 1993; Abend et al., 2010; Reichart et al., 2010b) as well as research on improving evaluation of unsupervised taggers (Frank et al., 2009; Reichart et al., 2010a).

Though they did not concentrate on unsupervised methods, Haghighi and Klein (2006) conducted an unsupervised experiment that utilized certain token features (e.g. character suffixes of 3 or less,

has initial capital, etc.; the features themselves are from Smith and Eisner (2005)) to learn parameters in an undirected graphical model which was the equivalent of an HMM in directed models. It was also the first study to posit the one-to-one evaluation criterion which has been repeated extensively since (Johnson, 2007; Headden III et al., 2008; Graca et al., 2009).

Finkel et al. (2007) is an interesting variant of unsupervised POS tagging where a parse tree is assumed and POS tags are induced from this structure non-parametrically. It is the converse of unsupervised parsing which assumes access to a tagged corpus and induces a parsing model.

Other models more directly influenced or closely parallel our work. Griffiths et al. (2005) is the work that inspired the current approach where a set of states is designated to capture variance across contexts. The primary goal of that model was to induce a topic model given data that had not been filtered of noise in the form of function words. As such, distinguishing between topic states such that they model different syntactic states was not attempted, and we have seen in sec. 3 that such an extension is not entirely straightforward.[4] Boyd-Graber and Blei (2009) has some parallels to our model in that a hidden variable over topics is distributed according to a normalized product between a context prior and a syntactic prior. However, it assumes a much greater amount of information than we do in that a parse tree as well as (possibly) POS tags are taken as observed. The model has a very different goal from ours as well, which is to infer a syntactically informed topic model. Teichert and Daumé III (2010) is another study with close similarities to our own. This study models distinctions between closed class words and open class words within a modified HMM. It is unclear from their formulation how the distinction between open class and closed class words is learned.

There is also extensive literature on learning sequence structure from *unlabeled* text (Smith and Eisner, 2005; Goldberg et al., 2008; Ravi and Knight, 2009) which assume access to a tag dictionary. Goldwater and Griffiths (2007) deserves mention for examining a semi-supervised model that sampled emission hyperparameters for each state rather than a single symmetric hyperparameter. They showed that this outperformed a symmetric model. An interesting heuristic model is Zhao and Marcus (2009) that uses a seed set of closed class words to classify open class words.

[4] We tested a variant of LDAHMM in which more than one state can generate topics. It did not achieve good results.

## 7 Conclusion

We have shown that a hidden Markov model that allocates a subset of the states to have distributions conditioned on localized domains can significantly improve performance in unsupervised part-of-speech tagging. We have also demonstrated that significant performance gains are possible simply by setting a different emission hyperparameter for a subgroup of the states. It is encouraging that these results hold for both models not just on the WSJ but across a diverse set of languages and measures.

We believe our proposed extensions to the HMM are a significant contribution to the general HMM and unsupervised POS tagging literature in that both can be implemented with minimum modification of existing MCMC inferred HMMs, have (nearly) equivalent run times, produce output that is easy to interpret since they are based on a generative framework, and bring about considerable performance improvements at the same time.

## Acknowledgments

The authors would like to thank Elias Ponvert and the anonymous reviewers. This work was supported by a grant from the Morris Memorial Trust Fund of the New York Community Trust.

## References

O. Abend, R. Reichart, and A. Rappoport. 2010. Improved unsupervised POS induction through prototype discovery. In *Proceedings of ACL*, pages 1298–1307.

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. Floresta sintá(c)tica": a treebank for Portuguese. In *Proceedings of LREC*, pages 1698–1703.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

J. L. Boyd-Graber and D. Blei. 2009. Syntactic topic models. In *Proceedings of NIPS*, pages 185–192.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

T. Brants. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of conference on Applied natural language processing*, pages 224–231.

A. Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*, pages 59–66.

J. R. Finkel, T. Grenager, and C. D. Manning. 2007. The infinite tree. In *Proceedings of ACL*, pages 272–279.

W.N. Francis, H. Kučera, and A.W. Mackie. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin Harcourt.

S. Frank, S. Goldwater, and F. Keller. 2009. Evaluating models of syntactic category acquisition without using a gold standard. In *Proceedings of CogSci*.

J. Gao and M. Johnson. 2008. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of EMNLP*, pages 344–352.

Y. Goldberg, M. Adler, and M. Elhadad. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of ACL*, pages 746–754.

S. Goldwater and T. L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, pages 744–751.

J. Graca, K. Ganchev, B. Taskar, and F. Pereira. 2009. Posterior vs parameter sparsity in latent variable models. In *Proceedings of NIPS*, pages 664–672.

T. L. Griffiths, M. Steyvers, D. M. Blei, and J. M. Tenenbaum. 2005. Integrating topics and syntax. In *Proceedings of NIPS*, pages 537–544.

A. Haghighi and D. Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of HLT/NAACL*, pages 320–327.

W. P. Headden III, D. McClosky, and E. Charniak. 2008. Evaluating unsupervised part-of-speech tagging for grammar induction. In *Proceedings of COLING*, pages 329–336.

G.E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

M. Johnson. 2007. Why doesn't EM find good HMM POS-taggers. In *Proceedings of EMNLP-CoNLL*, pages 296–305.

P. Liang and D. Klein. 2009. Online EM for unsupervised models. In *Proceedings of HLT/NAACL*, pages 611–619.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Comp. ling.*, 19(2):313–330.

M. Meila. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

A. Palmer, T. Moon, and J. Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL-HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44.

T. C. Pixabaj, M. A. Vicente Méndez, M. Vicente Méndez, and O. A. Damián. 2007. Text Collections in Four Mayan Languages. Archived in The Archive of the Indigenous Languages of Latin America.

S. Ravi and K. Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of ACL and AFNLP*, pages 504–512.

R. Reichart, O. Abend, and A. Rappoport. 2010a. Type level clustering evaluation: New measures and a POS induction case study. In *Proceedings of CoNLL*, pages 77–87.

R. Reichart, R. Fattal, and A. Rappoport. 2010b. Improved unsupervised POS induction using intrinsic clustering quality and a Zipfian constraint. In *Proceedings of CoNLL*, pages 57–66.

H. Schütze. 1993. Part-of-speech induction from scratch. In *Proceedings of ACL*, pages 251–258.

N.A. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of ACL*, pages 354–362.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of EMNLP*, pages 1041–1050.

A.R. Teichert and H. Daumé III. 2010. Unsupervised Part of Speech Tagging Without a Lexicon. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning 2010*.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL*, pages 173–180.

J. Van Gael, A. Vlachos, and Z. Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proceedings of EMNLP*, pages 678–687.

L. Xu, D. Wilkinson, F. Southey, and D. Schuurmans. 2006. Discriminative unsupervised learning of structured predictors. In *Proceedings of ICML*, pages 1057–1064.

Q. Zhao and M. Marcus. 2009. A simple unsupervised learner for POS disambiguation rules given only a minimal lexicon. In *Proceedings of EMNLP*, pages 688–697.