A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging*

Sharon Goldwater
Department of Linguistics
Stanford University
sgwater@stanford.edu

Thomas L. Griffiths
Department of Psychology
UC Berkeley
tom_griffiths@berkeley.edu

Abstract

Unsupervised learning of linguistic structure is a difficult problem. A common approach is to define a generative model and maximize the probability of the hidden structure given the observed data. Typically, this is done using maximum-likelihood estimation (MLE) of the model parameters. We show using part-of-speech tagging that a fully Bayesian approach can greatly improve performance. Rather than estimating a single set of parameters, the Bayesian approach integrates over all possible parameter values. This difference ensures that the learned structure will have high probability over a range of possible parameters, and permits the use of priors favoring the sparse distributions that are typical of natural language. Our model has the structure of a standard trigram HMM, yet its accuracy is closer to that of a state-of-the-art discriminative model (Smith and Eisner, 2005), up to 14 percentage points better than MLE. We find improvements both when training from data alone, and using a tagging dictionary.

1 Introduction

Unsupervised learning of linguistic structure is a difficult problem. Recently, several new model-based approaches have improved performance on a variety of tasks (Klein and Manning, 2002; Smith and Eisner, 2005). Nearly all of these approaches have one aspect in common: the goal of learning is to identify the set of model parameters that maximizes some objective function. Values for the hidden variables in the model are then chosen based on the learned parameterization. Here, we propose a different approach based on Bayesian statistical principles: rather than searching for an optimal set of parameter values, we seek to directly maximize the probability of the hidden variables given the observed data, integrating over all possible parameter values. Using part-of-speech (POS) tagging as an example application, we show that the Bayesian approach provides large performance improvements over maximum-likelihood estimation (MLE) for the same model structure. Two factors can explain the improvement. First, integrating over parameter values leads to greater robustness in the choice of tag sequence, since it must have high probability over a range of parameters. Second, integration permits the use of priors favoring sparse distributions, which are typical of natural language. These kinds of priors can lead to degenerate solutions if the parameters are estimated directly.

Before describing our approach in more detail, we briefly review previous work on unsupervised POS tagging. Perhaps the most well-known is that of Merialdo (1994), who used MLE to train a trigram hidden Markov model (HMM). More recent work has shown that improvements can be made by modifying the basic HMM structure (Banko and Moore, 2004), using better smoothing techniques or added constraints (Wang and Schuurmans, 2005), or using a discriminative model rather than an HMM

*This work was supported by grants NSF 0631518 and ONR MURI N000140510388. We would also like to thank Noah Smith for providing us with his data sets.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 744–751,
Prague, Czech Republic, June 2007.

(Smith and Eisner, 2005). Non-model-based approaches have also been proposed (Brill (1995); see also discussion in Banko and Moore (2004)). All of this work is really POS *disambiguation*: learning is strongly constrained by a dictionary listing the allowable tags for each word in the text. Smith and Eisner (2005) also present results using a diluted dictionary, where infrequent words may have any tag. Haghighi and Klein (2006) use a small list of labeled prototypes and no dictionary.

A different tradition treats the identification of syntactic classes as a knowledge-free clustering problem. Distributional clustering and dimensionality reduction techniques are typically applied when linguistically meaningful classes are desired (Schütze, 1995; Clark, 2000; Finch et al., 1995); probabilistic models have been used to find classes that can improve smoothing and reduce perplexity (Brown et al., 1992; Saul and Pereira, 1997). Unfortunately, due to a lack of standard and informative evaluation techniques, it is difficult to compare the effectiveness of different clustering methods.

In this paper, we hope to unify the problems of POS disambiguation and syntactic clustering by presenting results for conditions ranging from a full tag dictionary to no dictionary at all. We introduce the use of a new information-theoretic criterion, *variation of information* (Meilă, 2002), which can be used to compare a gold standard clustering to the clustering induced from a tagger's output, regardless of the cluster labels. We also evaluate using tag accuracy when possible. Our system outperforms an HMM trained with MLE on both metrics in all circumstances tested, often by a wide margin. Its accuracy in some cases is close to that of Smith and Eisner's (2005) discriminative model. Our results show that the Bayesian approach is particularly useful when learning is less constrained, either because less evidence is available (corpus size is small) or because the dictionary contains less information.

In the following section, we discuss the motivation for a Bayesian approach and present our model and search procedure. Section 3 gives results illustrating how the parameters of the prior affect results, and Section 4 describes how to infer a good choice of parameters from unlabeled data. Section 5 presents results for a range of corpus sizes and dictionary information, and Section 6 concludes.

2 A Bayesian HMM

2.1 Motivation

In model-based approaches to unsupervised language learning, the problem is formulated in terms of identifying latent structure from data. We define a model with parameters θ, some observed variables $\mathbf{w}$ (the linguistic input), and some latent variables $\mathbf{t}$ (the hidden structure). The goal is to assign appropriate values to the latent variables. Standard approaches do so by selecting values for the model parameters, and then choosing the most probable variable assignment based on those parameters. For example, maximum-likelihood estimation (MLE) seeks parameters $\hat{\theta}$ such that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\mathbf{w}|\theta), \tag{1}$$

where $P(\mathbf{w}|\theta) = \sum_{\mathbf{t}} P(\mathbf{w}, \mathbf{t}|\theta)$. Sometimes, a non-uniform prior distribution over θ is introduced, in which case $\hat{\theta}$ is the *maximum a posteriori* (MAP) solution for θ:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\mathbf{w}|\theta)P(\theta). \tag{2}$$

The values of the latent variables are then taken to be those that maximize $P(\mathbf{t}|\mathbf{w}, \hat{\theta})$.

In contrast, the Bayesian approach we advocate in this paper seeks to identify a distribution over latent variables directly, without ever fixing particular values for the model parameters. The distribution over latent variables given the observed data is obtained by integrating over all possible values of θ:

$$P(\mathbf{t}|\mathbf{w}) = \int P(\mathbf{t}|\mathbf{w}, \theta)P(\theta|\mathbf{w})d\theta. \tag{3}$$

This distribution can be used in various ways, including choosing the MAP assignment to the latent variables, or estimating expected values for them.

To see why integrating over possible parameter values can be useful when inducing latent structure, consider the following example. We are given a coin, which may be biased ($t = 1$) or fair ($t = 0$), each with probability .5. Let θ be the probability of heads. If the coin is biased, we assume a uniform distribution over θ, otherwise $\theta = .5$. We observe $\mathbf{w}$, the outcomes of 10 coin flips, and we wish to determine whether the coin is biased (i.e. the value of

t). Assume that we have a uniform prior on θ, with $p(\theta) = 1$ for all $\theta \in [0, 1]$. First, we apply the standard methodology of finding the MAP estimate for θ and then selecting the value of t that maximizes $P(t|\mathbf{w}, \hat{\theta})$. In this case, an elementary calculation shows that the MAP estimate is $\hat{\theta} = n_H/10$, where n_H is the number of heads in $\mathbf{w}$ (likewise, n_T is the number of tails). Consequently, $P(t|\mathbf{w}, \hat{\theta})$ favors $t = 1$ for any sequence that does not contain exactly five heads, and assigns equal probability to $t = 1$ and $t = 0$ for any sequence that does contain exactly five heads — a counterintuitive result. In contrast, using some standard results in Bayesian analysis we can show that applying Equation 3 yields

$$P(t = 1|\mathbf{w}) = 1 / \left(1 + \frac{11!}{n_H! n_T! 2^{10}}\right) \qquad (4)$$

which is significantly less than .5 when $n_H = 5$, and only favors $t = 1$ for sequences where $n_H \geq 8$ or $n_H \leq 2$. This intuitively sensible prediction results from the fact that the Bayesian approach is sensitive to the *robustness* of a choice of t to the value of θ, as illustrated in Figure 1. Even though a sequence with $n_H = 6$ yields a MAP estimate of $\hat{\theta} = 0.6$ (Figure 1 (a)), $P(t = 1|\mathbf{w}, \theta)$ is only greater than 0.5 for a small range of θ around $\hat{\theta}$ (Figure 1 (b)), meaning that the choice of $t = 1$ is not very robust to variation in θ. In contrast, a sequence with $n_H = 8$ favors $t = 1$ for a wide range of θ around $\hat{\theta}$. By integrating over θ, Equation 3 takes into account the consequences of possible variation in θ.

Another advantage of integrating over θ is that it permits the use of linguistically appropriate priors. In many linguistic models, including HMMs, the distributions over variables are multinomial. For a multinomial with parameters $\theta = (\theta_1, \ldots, \theta_K)$, a natural choice of prior is the K-dimensional Dirichlet distribution, which is conjugate to the multinomial.[1] For simplicity, we initially assume that all K parameters (also known as *hyperparameters*) of the Dirichlet distribution are equal to β, i.e. the Dirichlet is *symmetric*. The value of β determines which parameters θ will have high probability: when $\beta = 1$, all parameter values are equally likely; when $\beta > 1$, multinomials that are closer to uniform are preferred; and when $\beta < 1$, high probability is assigned to sparse multinomials, where one or more parameters are at or near 0.

[1]A prior is *conjugate* to a distribution if the posterior has the same form as the prior.

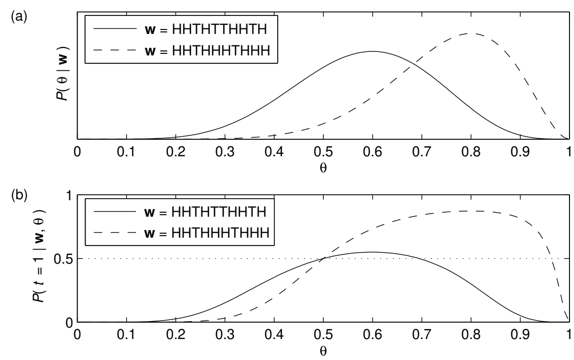

Figure 1: The Bayesian approach to estimating the value of a latent variable, t, from observed data, $\mathbf{w}$, chooses a value of t robust to uncertainty in θ. (a) Posterior distribution on θ given $\mathbf{w}$. (b) Probability that $t = 1$ given $\mathbf{w}$ and θ as a function of θ.

Typically, linguistic structures are characterized by sparse distributions (e.g., POS tags are followed with high probability by only a few other tags, and have highly skewed output distributions). Consequently, it makes sense to use a Dirichlet prior with $\beta < 1$. However, as noted by Johnson et al. (2007), this choice of β leads to difficulties with MAP estimation. For a sequence of draws $\mathbf{x} = (x_1, \ldots, x_n)$ from a multinomial distribution θ with observed counts $n_1, \ldots, n_K$, a symmetric Dirichlet(β) prior over θ yields the MAP estimate $\theta_k = \frac{n_k + \beta - 1}{n + K(\beta - 1)}$. When $\beta \geq 1$, standard MLE techniques such as EM can be used to find the MAP estimate simply by adding "pseudocounts" of size $\beta - 1$ to each of the expected counts n_k at each iteration. However, when $\beta < 1$, the values of θ that set one or more of the θ_k equal to 0 can have infinitely high posterior probability, meaning that MAP estimation can yield degenerate solutions. If, instead of estimating θ, we integrate over all possible values, we no longer encounter such difficulties. Instead, the probability that outcome x_i takes value k given previous outcomes $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1})$ is

$$\begin{aligned} P(k|\mathbf{x}_{-i}, \beta) &= \int P(k|\theta) P(\theta|\mathbf{x}_{-i}, \beta)\, d\theta \\ &= \frac{n_k + \beta}{i - 1 + K\beta} \qquad (5) \end{aligned}$$

where n_k is the number of times k occurred in $\mathbf{x}_{-i}$. See MacKay and Peto (1995) for a derivation.

2.2 Model Definition

Our model has the structure of a standard trigram HMM, with the addition of symmetric Dirichlet priors over the transition and output distributions:

$$
\begin{array}{ll}
t_i | t_{i-1} = t, t_{i-2} = t', \tau^{(t,t')} & \sim \text{Mult}(\tau^{(t,t')}) \\
w_i | t_i = t, \omega^{(t)} & \sim \text{Mult}(\omega^{(t)}) \\
\tau^{(t,t')} | \alpha & \sim \text{Dirichlet}(\alpha) \\
\omega^{(t)} | \beta & \sim \text{Dirichlet}(\beta)
\end{array}
$$

where t_i and w_i are the ith tag and word. We assume that sentence boundaries are marked with a distinguished tag. For a model with T possible tags, each of the transition distributions $\tau^{(t,t')}$ has T components, and each of the output distributions $\omega^{(t)}$ has W_t components, where W_t is the number of word types that are permissible outputs for tag t. We will use τ and ω to refer to the entire transition and output parameter sets. This model assumes that the prior over state transitions is the same for all histories, and the prior over output distributions is the same for all states. We relax the latter assumption in Section 4.

Under this model, Equation 5 gives us

$$P(t_i | \mathbf{t}_{-i}, \alpha) = \frac{n_{(t_{i-2}, t_{i-1}, t_i)} + \alpha}{n_{(t_{i-2}, t_{i-1})} + T\alpha} \quad (6)$$

$$P(w_i | t_i, \mathbf{t}_{-i}, \mathbf{w}_{-i}, \beta) = \frac{n_{(t_i, w_i)} + \beta}{n_{(t_i)} + W_{t_i}\beta} \quad (7)$$

where $n_{(t_{i-2}, t_{i-1}, t_i)}$ and $n_{(t_i, w_i)}$ are the number of occurrences of the trigram (t_{i-2}, t_{i-1}, t_i) and the tag-word pair (t_i, w_i) in the $i-1$ previously generated tags and words. Note that, by integrating out the parameters τ and ω, we induce dependencies between the variables in the model. The probability of generating a particular trigram tag sequence (likewise, output) depends on the number of times that sequence (output) has been generated previously. Importantly, trigrams (and outputs) remain *exchangeable*: the probability of a set of trigrams (outputs) is the same regardless of the order in which it was generated. The property of exchangeability is crucial to the inference algorithm we describe next.

2.3 Inference

To perform inference in our model, we use Gibbs sampling (Geman and Geman, 1984), a stochastic procedure that produces samples from the posterior distribution $P(\mathbf{t}|\mathbf{w}, \alpha, \beta) \propto P(\mathbf{w}|\mathbf{t}, \beta)P(\mathbf{t}|\alpha)$. We initialize the tags at random, then iteratively resample each tag according to its conditional distribution given the current values of all other tags. Exchangeability allows us to treat the current counts of the other tag trigrams and outputs as "previous" observations. The only complication is that resampling a tag changes the identity of three trigrams at once, and we must account for this in computing its conditional distribution. The sampling distribution for t_i is given in Figure 2.

In Bayesian statistical inference, multiple samples from the posterior are often used in order to obtain statistics such as the expected values of model variables. For POS tagging, estimates based on multiple samples might be useful if we were interested in, for example, the probability that two words have the same tag. However, computing such probabilities across all pairs of words does not necessarily lead to a consistent clustering, and the result would be difficult to evaluate. Using a single sample makes standard evaluation methods possible, but yields suboptimal results because the value for each tag is sampled from a distribution, and some tags will be assigned low-probability values. Our solution is to treat the Gibbs sampler as a stochastic search procedure with the goal of identifying the MAP tag sequence. This can be done using tempering (annealing), where a temperature of ϕ is equivalent to raising the probabilities in the sampling distribution to the power of $\frac{1}{\phi}$. As ϕ approaches 0, even a single sample will provide a good MAP estimate.

3 Fixed Hyperparameter Experiments

3.1 Method

Our initial experiments follow in the tradition begun by Merialdo (1994), using a tag dictionary to constrain the possible parts of speech allowed for each word. (This also fixes W_t, the number of possible words for tag t.) The dictionary was constructed by listing, for each word, all tags found for that word in the entire WSJ treebank. For the experiments in this section, we used a 24,000-word subset of the tree-

$$P(t_i|\mathbf{t}_{-i},\mathbf{w},\alpha,\beta) \propto \frac{n_{(t_i,w_i)}+\beta}{n_{t_i}+W_{t_i}\beta}\cdot\frac{n_{(t_{i-2},t_{i-1},t_i)}+\alpha}{n_{(t_{i-2},t_{i-1})}+T\alpha}\cdot\frac{n_{(t_{i-1},t_i,t_{i+1})}+I(t_{i-2}=t_{i-1}=t_i=t_{i+1})+\alpha}{n_{(t_{i-1},t_i)}+I(t_{i-2}=t_{i-1}=t_i)+T\alpha}$$
$$\cdot\frac{n_{(t_i,t_{i+1},t_{i+2})}+I(t_{i-2}=t_i=t_{i+2},t_{i-1}=t_{i+1})+I(t_{i-1}=t_i=t_{i+1}=t_{i+2})+\alpha}{n_{(t_i,t_{i+1})}+I(t_{i-2}=t_i,t_{i-1}=t_{i+1})+I(t_{i-1}=t_i=t_{i+1})+T\alpha}$$

Figure 2: Conditional distribution for t_i. Here, $\mathbf{t}_{-i}$ refers to the current values of all tags except for t_i, $I(.)$ is a function that takes on the value 1 when its argument is true and 0 otherwise, and all counts n_x are with respect to the tag trigrams and tag-word pairs in $(\mathbf{t}_{-i},\mathbf{w}_{-i})$.

bank as our unlabeled training corpus. 54.5% of the tokens in this corpus have at least two possible tags, with the average number of tags per token being 2.3. We varied the values of the hyperparameters α and β and evaluated overall tagging accuracy. For comparison with our Bayesian HMM (BHMM) in this and following sections, we also present results from the Viterbi decoding of an HMM trained using MLE by running EM to convergence (MLHMM). Where direct comparison is possible, we list the scores reported by Smith and Eisner (2005) for their conditional random field model trained using contrastive estimation (CRF/CE).[2]

For all experiments, we ran our Gibbs sampling algorithm for 20,000 iterations over the entire data set. The algorithm was initialized with a random tag assignment and a temperature of 2, and the temperature was gradually decreased to .08. Since our inference procedure is stochastic, our reported results are an average over 5 independent runs.

Results from our model for a range of hyperparameters are presented in Table 1. With the best choice of hyperparameters ($\alpha = .003, \beta = 1$), we achieve average tagging accuracy of 86.8%. This far surpasses the MLHMM performance of 74.5%, and is closer to the 90.1% accuracy of CRF/CE on the same data set using oracle parameter selection. The effects of α, which determines the probability of the transition distributions, are stronger than the effects of β, which determines the probability of the output distributions. The optimal value of .003 for α reflects the fact that the true transition probability matrix for this corpus is indeed sparse. As α grows larger, the model prefers more uniform transition probabilities, which causes it to perform worse. Although the true output distributions tend to be sparse as well, the level of sparseness depends on the tag (consider function words vs. content words in particular). Therefore, a value of β that accurately reflects the most probable output distributions for some tags may be a poor choice for other tags. This leads to the smaller effect of β, and suggests that performance might be improved by selecting a different β for each tag, as we do in the next section.

Value of α	Value of β .001	.003	.01	.03	.1	.3	1.0
.001	85.0	85.7	86.1	86.0	86.2	86.5	86.6
.003	85.5	85.5	85.8	86.6	86.7	86.7	**86.8**
.01	85.3	85.5	85.6	85.9	86.4	86.4	86.2
.03	85.9	85.8	86.1	86.2	86.6	86.8	86.4
.1	85.2	85.0	85.2	85.1	84.9	85.5	84.9
.3	84.4	84.4	84.6	84.4	84.5	85.7	85.3
1.0	83.1	83.0	83.2	83.3	83.5	83.7	83.9

Table 1: Percentage of words tagged correctly by BHMM as a function of the hyperparameters α and β. Results are averaged over 5 runs on the 24k corpus with full tag dictionary. Standard deviations in most cases are less than .5.

A final point worth noting is that even when $\alpha = \beta = 1$ (i.e., the Dirichlet priors exert no influence) the BHMM still performs much better than the MLHMM. This result underscores the importance of integrating over model parameters: the BHMM identifies a sequence of tags that have high proba-

[2]Results of CRF/CE depend on the set of features used and the contrast neighborhood. In all cases, we list the best score reported for any contrast neighborhood using trigram (but no spelling) features. To ensure proper comparison, all corpora used in our experiments consist of the same randomized sets of sentences used by Smith and Eisner. Note that training on sets of contiguous sentences from the beginning of the treebank consistently improves our results, often by 1-2 percentage points or more. MLHMM scores show less difference between randomized and contiguous corpora.

bility over a range of parameter values, rather than choosing tags based on the single best set of parameters. The improved results of the BHMM demonstrate that selecting a sequence that is robust to variations in the parameters leads to better performance.

4 Hyperparameter Inference

In our initial experiments, we experimented with different fixed values of the hyperparameters and reported results based on their optimal values. However, choosing hyperparameters in this way is time-consuming at best and impossible at worst, if there is no gold standard available. Luckily, the Bayesian approach allows us to automatically select values for the hyperparameters by treating them as additional variables in the model. We augment the model with priors over the hyperparameters (here, we assume an improper uniform prior), and use a single Metropolis-Hastings update (Gilks et al., 1996) to resample the value of each hyperparameter after each iteration of the Gibbs sampler. Informally, to update the value of hyperparameter α, we sample a proposed new value α' from a normal distribution with $\mu = \alpha$ and $\sigma = .1\alpha$. The probability of accepting the new value depends on the ratio between $P(\mathbf{t}|\mathbf{w}, \alpha)$ and $P(\mathbf{t}|\mathbf{w}, \alpha')$ and a term correcting for the asymmetric proposal distribution.

Performing inference on the hyperparameters allows us to relax the assumption that every tag has the same prior on its output distribution. In the experiments reported in the following section, we used two different versions of our model. The first version (BHMM1) uses a single value of β for all word classes (as above); the second version (BHMM2) uses a separate β_j for each tag class j.

5 Inferred Hyperparameter Experiments

5.1 Varying corpus size

In this set of experiments, we used the full tag dictionary (as above), but performed inference on the hyperparameters. Following Smith and Eisner (2005), we trained on four different corpora, consisting of the first 12k, 24k, 48k, and 96k words of the WSJ corpus. For all corpora, the percentage of ambiguous tokens is 54%-55% and the average number of tags per token is 2.3. Table 2 shows results for the various models and a random baseline (averaged over 5 random tag assignments). Hyperparameter inference leads to slightly lower scores than are obtained by oracle hyperparameter selection, but both versions of BHMM are still far superior to MLHMM for all corpus sizes. Not surprisingly, the advantages of BHMM are most pronounced on the smallest corpus: the effects of parameter integration and sensible priors are stronger when less evidence is available from the input. In the limit as corpus size goes to infinity, the BHMM and MLHMM will make identical predictions.

	Corpus size			
Accuracy	12k	24k	48k	96k
random	64.8	64.6	64.6	64.6
MLHMM	71.3	74.5	76.7	78.3
CRF/CE	86.2	88.6	88.4	89.4
BHMM1	85.8	85.2	83.6	85.0
BHMM2	85.8	84.4	85.7	85.8
$\sigma <$	.7	.2	.6	.2

Table 2: Percentage of words tagged correctly by the various models on different sized corpora. BHMM1 and BHMM2 use hyperparameter inference; CRF/CE uses parameter selection based on an unlabeled development set. Standard deviations (σ) for the BHMM results fell below those shown for each corpus size.

5.2 Varying dictionary knowledge

In unsupervised learning, it is not always reasonable to assume that a large tag dictionary is available. To determine the effects of reduced or absent dictionary information, we ran a set of experiments inspired by those of Smith and Eisner (2005). First, we collapsed the set of 45 treebank tags onto a smaller set of 17 (the same set used by Smith and Eisner). We created a full tag dictionary for this set of tags from the entire treebank, and also created several reduced dictionaries. Each reduced dictionary contains the tag information only for words that appear at least d times in the training corpus (the 24k corpus, for these experiments). All other words are fully ambiguous between all 17 classes. We ran tests with $d = 1, 2, 3, 5, 10$, and ∞ (i.e., knowledge-free syntactic clustering).

With standard accuracy measures, it is difficult to

	Value of d					
Accuracy	1	2	3	5	10	∞
random	69.6	56.7	51.0	45.2	38.6	
MLHMM	83.2	70.6	65.5	59.0	50.9	
CRF/CE	90.4	77.0	71.7			
BHMM1	86.0	76.4	71.0	64.3	58.0	
BHMM2	87.3	79.6	65.0	59.2	49.7	
$\sigma <$	.2	.8	.6	.3	1.4	
VI						
random	2.65	3.96	4.38	4.75	5.13	7.29
MLHMM	1.13	2.51	3.00	3.41	3.89	6.50
BHMM1	1.09	2.44	2.82	3.19	3.47	4.30
BHMM2	**1.04**	**1.78**	**2.31**	**2.49**	**2.97**	**4.04**
$\sigma <$	.02	.03	.04	.03	.07	.17
Corpus stats						
% ambig.	49.0	61.3	66.3	70.9	75.8	100
tags/token	1.9	4.4	5.5	6.8	8.3	17

Table 3: Percentage of words tagged correctly and variation of information between clusterings induced by the assigned and gold standard tags as the amount of information in the dictionary is varied. Standard deviations (σ) for the BHMM results fell below those shown in each column. The percentage of ambiguous tokens and average number of tags per token for each value of d is also shown.

evaluate the quality of a syntactic clustering when no dictionary is used, since cluster names are interchangeable. We therefore introduce another evaluation measure for these experiments, a distance metric on clusterings known as *variation of information* (Meilă, 2002). The variation of information (VI) between two clusterings C (the gold standard) and C' (the found clustering) of a set of data points is a sum of the amount of information lost in moving from C to C', and the amount that must be gained. It is defined in terms of entropy H and mutual information I: $VI(C, C') = H(C) + H(C') - 2I(C, C')$. Even when accuracy can be measured, VI may be more informative: two different tag assignments may have the same accuracy but different VI with respect to the gold standard if the *errors* in one assignment are less consistent than those in the other.

Table 3 gives the results for this set of experiments. One or both versions of BHMM outperform MLHMM in terms of tag accuracy for all values of d, although the differences are not as great as in earlier experiments. The differences in VI are more striking, particularly as the amount of dictionary information is reduced. When ambiguity is greater, both versions of BHMM show less confusion with respect to the true tags than does MLHMM, and BHMM2 performs the best in all circumstances. The confusion matrices in Figure 3 provide a more intuitive picture of the very different sorts of clusterings produced by MLHMM and BHMM2 when no tag dictionary is available. Similar differences hold to a lesser degree when a partial dictionary is provided. With MLHMM, different tokens of the same word type are usually assigned to the same cluster, but types are assigned to clusters more or less at random, and all clusters have approximately the same number of types (542 on average, with a standard deviation of 174). The clusters found by BHMM2 tend to be more coherent and more variable in size: in the 5 runs of BHMM2, the average number of types per cluster ranged from 436 to 465 (i.e., tokens of the same word are spread over fewer clusters than in MLHMM), with a standard deviation between 460 and 674. Determiners, prepositions, the possessive marker, and various kinds of punctuation are mostly clustered coherently. Nouns are spread over a few clusters, partly due to a distinction found between common and proper nouns. Likewise, modal verbs and the copula are mostly separated from other verbs. Errors are often sensible: adjectives and nouns are frequently confused, as are verbs and adverbs.

The kinds of results produced by BHMM1 and BHMM2 are more similar to each other than to the results of MLHMM, but the differences are still informative. Recall that BHMM1 learns a single value for β that is used for all output distributions, while BHMM2 learns separate hyperparameters for each cluster. This leads to different treatments of difficult-to-classify low-frequency items. In BHMM1, these items tend to be spread evenly among all clusters, so that all clusters have similarly sparse output distributions. In BHMM2, the system creates one or two clusters consisting entirely of very infrequent items, where the priors on these clusters strongly prefer uniform outputs, and all other clusters prefer extremely sparse outputs (and are more coherent than in BHMM1). This explains the difference in VI between the two systems, as well as the higher accuracy of BHMM1 for $d \geq 3$: the single β discourages placing low-frequency items in their own cluster, so they are more likely to be clustered with items that have sim-

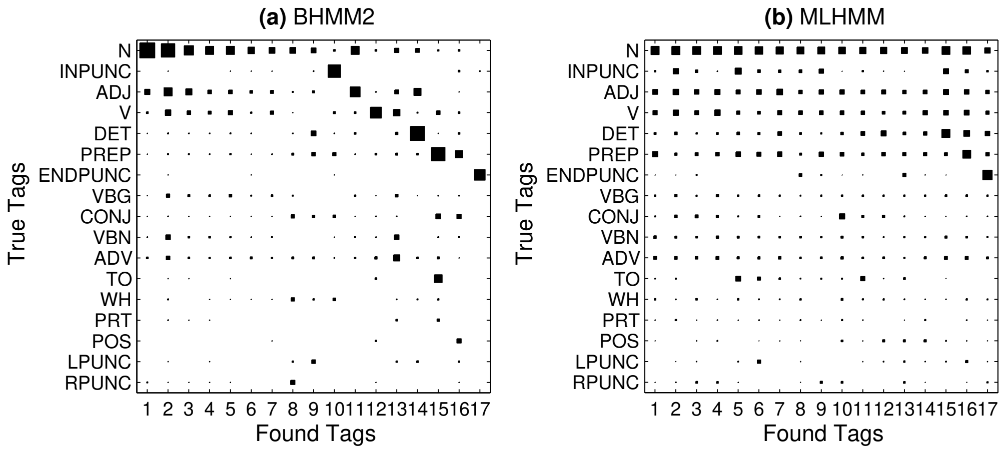

Figure 3: Confusion matrices for the dictionary-free clusterings found by (a) BHMM2 and (b) MLHMM.

ilar transition probabilities. The problem of junk clusters in BHMM2 might be alleviated by using a non-uniform prior over the hyperparameters to encourage some degree of sparsity in all clusters.

6 Conclusion

In this paper, we have demonstrated that, for a standard trigram HMM, taking a Bayesian approach to POS tagging dramatically improves performance over maximum-likelihood estimation. Integrating over possible parameter values leads to more robust solutions and allows the use of priors favoring sparse distributions. The Bayesian approach is particularly helpful when learning is less constrained, either because less data is available or because dictionary information is limited or absent. For knowledge-free clustering, our approach can also be extended through the use of infinite models so that the number of clusters need not be specified in advance. We hope that our success with POS tagging will inspire further research into Bayesian methods for other natural language learning tasks.

References

M. Banko and R. Moore. 2004. A study of unsupervised part-of-speech tagging. In *Proceedings of COLING '04*.

E. Brill. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 1–13.

P. Brown, V. Della Pietra, V. de Souza, J. Lai, and R. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

A. Clark. 2000. Inducing syntactic categories by context distribution clustering. In *Proceedings of the Conference on Natural Language Learning (CONLL)*.

S. Finch, N. Chater, and M. Redington. 1995. Acquiring syntactic information from distributional statistics. In J. In Levy, D. Bairaktaris, J. Bullinaria, and P. Cairns, editors, *Connectionist Models of Memory and Language*. UCL Press, London.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

W.R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Suffolk.

A. Haghighi and D. Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of HLT-NAACL*.

M. Johnson, T. Griffiths, and S. Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo.

D. Klein and C. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the ACL*.

D. MacKay and L. Bauman Peto. 1995. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1:289–307.

M. Meilă. 2002. Comparing clusterings. Technical Report 418, University of Washington Statistics Department.

B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

L. Saul and F. Pereira. 1997. Aggregate and mixed-order markov models for statistical language processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

H. Schütze. 1995. Distributional part-of-speech tagging. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

N. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of ACL*.

I. Wang and D. Schuurmans. 2005. Improved estimation for unsupervised part-of-speech tagging. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE)*.