# Improved Estimation for Unsupervised Part-of-Speech Tagging

Qin Iris Wang Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
{wqin,dale}@cs.ualberta.ca

***Abstract*— We demonstrate that a simple hidden Markov model can achieve state of the art performance in unsupervised part-of-speech tagging, by improving aspects of standard Baum-Welch (EM) estimation. One improvement uses word similarities to smooth the lexical *tag→word* probability estimates, which avoids over-fitting the lexical model. Another improvement constrains the model to preserve a specified marginal distribution over the hidden tags, which avoids over-fitting the *tag→tag* transition model. Although using more contextual information than an HMM remains desirable, improving basic estimation still leads to significant improvements and remains a prerequisite for training more complex models.**

## I. Introduction

A recent trend in statistical language learning research is to develop models that consider greater context when inferring latent linguistic variables such as part-of-speech tags, parsing dependencies, modifier attachments and constituents. This emphasis is exemplified by recent work on maximum entropy methods [1], [2], conditional random fields [3], and large margin methods [4], which often consider longer-range features than local words or n-grams. Even in the specific case of part-of-speech tagging, there exists a similar trend toward using greater context [5].

However, in this paper we present an alternative viewpoint by considering only simple models that use limited context. Our goal is to focus on the parameter estimation problem and show that simple ideas can still lead to significant improvements in training accuracy—ultimately leading to performance that competes with that of more complex models.

Although building complex models that capture wider linguistic structure is a legitimate goal of statistical language learning, ensuring the accuracy of these models becomes more difficult with increasing complexity. With a greater number of parameters, the computational challenges associated with training become more difficult, while the risks of over-fitting the training data simultaneously increase. Our hope is that improvements in basic estimation for simple models can be transferred to more complex settings.

In this paper, we focus on *unsupervised* language learning. Unsupervised learning is clearly important in statistical natural language research as it eliminates the need for extensive manual annotation. However, unsupervised learning is significantly harder than supervised learning, especially on language data where so many linguistic variables remain latent. The dangers of poor estimation are exaggerated and even self-reinforcing in this case, since unsupervised learners typically attempt to bootstrap from inferred values of hidden variables, which themselves are based on earlier, weaker estimates.

To address the issue of improving estimation in unsupervised learning, we investigate two simple ideas for improving the quality of *hidden Markov model* (HMM) training for unsupervised part-of-speech tagging. The first idea we present is to add a constraint that appropriate marginal tag probabilities be preserved. The second idea is to smooth the lexical parameters using word similarities. Each idea on its own achieves state of the art performance for unsupervised tagging.

Before presenting our results in more detail, we briefly review work on unsupervised part-of-speech tagging, introduce the HMM approach we use, and highlight the key shortcomings of the standard (EM) training procedure for this problem.

## II. Unsupervised part-of-speech Tagging

Automated part-of-speech tagging has been extensively investigated for more than a decade. However, most research has focused on *supervised* tagging where the availability of manually tagged data is assumed. Nevertheless, research on supervised tagging remains active, motivated in part by recent developments in maximum entropy estimation [1], conditional random fields [3], and large margin methods [4]. By comparison, work on unsupervised part-of-speech tagging has been less prevalent. A notable exception is the recent work [6], [7] which demonstrates renewed interest in unsupervised tagging.

Most research on unsupervised part-of-speech tagging was performed in the early 1990's. Consistently, the dominant approaches have been based on HMMs [8], [9], [10], [11], [12], with the notable exception of the transformation-based learning approach of [13]. Previous work has primarily focused on extending HMMs to incorporate more context [6], [8], [9], although some work has also considered techniques for improving parameter estimation quality [6], [7], [8], [10], [11], [12]. Our focus in this paper is on improving estimation quality rather than extending the model. In fact, we will simply consider a standard HMM, which is a simpler model than generally considered in previous work.

Recently, Banko and Moore [6] made the startling observation that much of the success of previous unsupervised taggers was due to the use of artificially *reduced* lexicons.

That is, researchers had manipulated lexicons by reducing the set of possible tags for each word to just those tags that have sufficiently high conditional probability. Banko and Moore [6] observe that if the true lexicon were used (consisting of the the true set of legal tags for each word) instead of the artificially ablated lexicon (consisting only of high probability tags for each word), the word tagging error rates of these unsupervised taggers skyrockets more than five-fold—from about 4.1% error to 22.8% error.

Unfortunately, as Banko and Moore [6] point out, the reliance on artificially reduced lexicons eliminates much of the unsupervised nature of the process, since the conditional probabilities are obtained from tagged data generally. In fact, this form of lexicon editing removes linguistic knowledge that is essentially correct and easily obtainable (from dictionaries), and replaces it with information that requires tagged data and/or manual editing to obtain. One goal of this paper is to show that artificial lexicon restriction might not be ultimately necessary. In particular, we hope to obtain accurate unsupervised tagging performance using the standard, full lexicon.

Currently, the best unsupervised tagging performance achieved with the full, unedited lexicon (from the complete Treebank) is an unimpressive 22.8% word tagging error rate [6]. Below we will see that a few simple improvements to the EM training procedure can reduce this error to 9.5%, while still using the full lexicon. Our new result appears to be closing in on previous error rates that were only obtained using artificially reduced lexicons.

To ensure our results are replicable, we follow [6], [5], [4] and use the Penn Treebank 3 corpus. We use sections 00-18 for training, 19-21 for development, and 22-24 for testing. Also following Banko and Moore [6] we use the full lexicon obtained from the full set of tags that occur for each word in the Treebank (but use no other information about the conditional probabilities of tags given words). When we use a reduced lexicon for comparison purposes, we remove the tags that occur with probability less than 0.1 for each word. Any other data resources we used we mention explicitly below.

## III. Unsupervised HMM tagging

Our investigation is based on the standard HMM which forms the foundation for most work in this area. An HMM specifies a joint probability distribution over a word and tag sequence, $\mathbf{w} = (w_1, ..., w_n)$ and $\mathbf{t} = (t_1, ..., t_n)$, where each word $w_i$ is assumed to be conditionally independent of the remaining words and tags given its part-of-speech tag $t_i$, and subsequent part-of-speech tags $t_{i+k}$ are assumed to be conditionally independent of previous tags $t_{i-k}$ given $t_i$.

The parameters of an HMM are given by *tag→tag* transition probabilities

$$a_{st} = P(T_{i+1} = t | T_i = s)$$

and *tag→word* emission probabilities

$$b_{tw} = P(W_i = w | T_i = t) \ ^1$$

If these parameters are known, then various queries can be computed efficiently, such as the probability of a joint tag/word sequence

$$P(\mathbf{tw}) = \prod_{i=1}^{n} a_{t_{i-1}t_i} b_{t_i w_i}.$$

More interestingly, the marginal probabilities of a tag and adjacent pair of tags, given the word sequence, $P(t_i|\mathbf{w})$ and $P(t_i t_{i+1}|\mathbf{w})$, can both be efficiently computed using the *forward-backward* algorithm [14]. Importantly, in our case, tagging can be accomplished by taking an untagged word sequence $\mathbf{w}$, and then computing the optimal tagging under the assumed probability model

$$\mathbf{t}^* = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{w})$$

where the optimal tag sequence can be efficiently recovered using the *Viterbi* algorithm.

It is apparent that, as a tagger, the HMM uses limited context. That is, the tag $t_i^*$ assigned to word $w_i$ depends directly only on $t_{i-1}$, $w_i$ and $t_{i+1}$, and is independent of the rest of the model given these values. As many researchers have pointed out, from a linguistic perspective, one would expect a much larger context to be relevant to determining the tag of word $w_i$ [5], [8]. Nevertheless, even for the simple HMM model which employs limited context, we will see that if appropriate values can be obtained for the parameters, the resulting tagger outperforms current unsupervised taggers (using the full lexicon).

Our focus is on the estimation problem: how can HMM parameters be estimated from untagged sentences $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ...$ while still obtaining an accurate tagger? The classical approach is simply to use expectation-maximization (EM) to find the transition and emission parameters $\mathbf{a}$ and $\mathbf{b}$ that locally maximize the marginal loglikelihood of the observed data

$$\sum_j \log P(\mathbf{w}^{(j)} | \mathbf{a}, \mathbf{b})$$

For HMMs this is often referred to as Baum-Welch training. Although EM is guaranteed to converge to a local maximum of this objective [15], the marginal log likelihood for an HMM tends to have multiple local maxima. Consequently, the resulting quality of the parameters can have more to do with initialization than the EM procedure itself. Another weakness of EM is that it is not informed by any particular linguistic knowledge. Although the training data provides guidance, and the model parameters might be linguistically meaningful, there are basic properties of these parameters that EM cannot readily discover on its own (which we will observe below), even given large amounts of unlabeled data.

### A. Baseline Experiment

To provide a baseline performance measure, as well as highlight some of the key difficulties, we first consider the results of training an HMM using standard Baum-Welch (EM) estimation. Initialization is crucial. Here we initialize the *tag→tag* transition parameters to be uniform over all tags, and the *tag→word* emission parameters as uniform over the

[1] Note that we assume a special start word $w_0$ and start tag $t_0$ occurs at the beginning of every sentence, which means we do not require additional parameters describing the initial tag distribution $P(t_1)$.

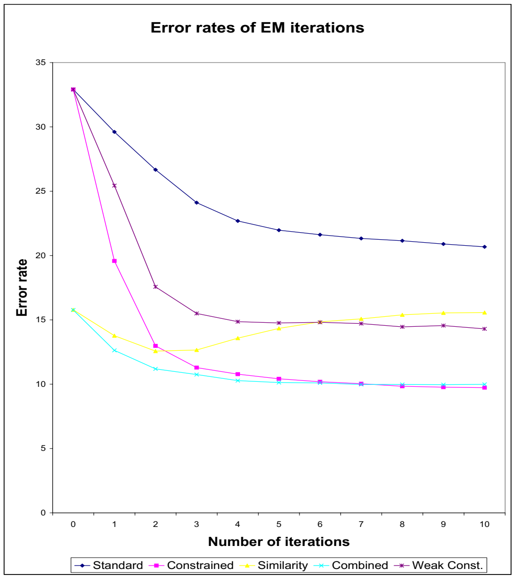

Fig. 1. Tagging error rate versus EM iteration for various HMM training techniques.

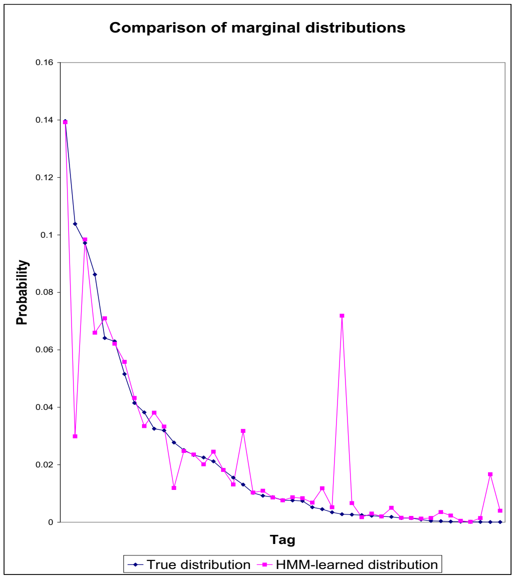

Fig. 2. Comparison of marginal tag probability distributions. X-axis means 45 different tags.

set of all possible words for a given tag (as specified in the complete lexicon).

Figure 1 shows that the initial model ("Standard") obtains a word-tag error rate of 34.1% on the test data, which progressively improves to 18.7% error after several EM iterations. (See also Table I.) Surprisingly, this preliminary result already demonstrates a smaller error than the 22.8% error reported in [6], even though we are using a simpler HMM model incorporating less context. However, it is unclear how many EM iterations were used in [6]. Despite this discrepancy, the results remain roughly in the same ballpark. Although 18.7% error already appears to be the best known result for unsupervised part-of-speech taggers trained with the full lexicon, it is still substantially larger than the 4.1% error obtained by Banko and Moore [6] using the artificially ablated lexicon. (For the record, using the ablated lexicon, the standard HMM initialized and trained as above obtains 6.2% error.)

Rather than resort to lexicon ablation, we can instead analyze the outcome of HMM training and attempt to identify improvements to the estimation procedure that might reduce error while still using the full lexicon.

Recall that there are two separate components of the HMM being estimated, the transition model and the emission model. Each component has its own weakness. First, for the learned transition model, we find that the estimated parameters are actually quite poor. Figure 2 shows a comparison of the raw tag frequencies produced by the learned HMM tagger, versus the true tag frequencies over the complete Treebank. Quite clearly these two distributions differ significantly. For example, the LS tag is predicted 72 times more often by the HMM than its true frequency in the corpus. (LS denotes a list item label, such as 1, 2, 3, or a, b, c.) Further investigation reveals that the word ”a” is always tagged as LS by the HMM, as opposed to the much more likely DT tag (determiner). This phenomenon was also observed by Banko and Moore [6]. In effect, the HMM has "learned" that the LS tag denotes another form of determiner. Although this is clearly a mistake, the question is what aspect of EM might prevent such an error for occurring? For the standard Baum-Welch algorithm it turns out that the answer is: nothing. Our first idea below addresses this shortcoming.

The second component of the estimated HMM, the lexical *tag*→*word* probabilities, are also quite poor. For example, the two prepositions "from" and "at" have very similar behavior under the true distribution (e.g. $P(\text{"from"}|\text{IN}) = 0.0442$ and $P(\text{"at"}|\text{IN}) = 0.0438$) and yet the parameter estimates produced by EM are widely separated in this case (e.g. $\hat{P}(\text{"from"}|\text{IN}) = 0.1488$ versus $\hat{P}(\text{"at"}|\text{IN}) = 0.0112$). A key weakness of EM is that it incorporates no form of "parameter tying" over the word emission parameters. That is, the distinct word emission parameters, $b_{tw_1}$ and $b_{tw_2}$, are treated independently by the optimization while it maximizes likelihood. However, this does not directly enforce the fact that many words play similar syntactic roles and should therefore have similar parameter values. EM can only exert parameter tying indirectly through the likelihood objective, but this constraint turns out to be too weak, as most words occur rarely. Consequently, EM tends to over-fit the lexical model and learn distinct parameter values for similar words.

This problem has been recognized in past research [8] and has been addressed by clustering rare words into equivalence classes. However, similar words rarely behave identically, and a simple clustering approach can actually over-smooth the estimates. As an alternative, we propose to use a natural form of *word similarity* to smooth these estimates more effectively. This is the second idea we consider below.

Note that extending the HMM approach to incorporate more context is desirable from a modeling perspective. However, adding complexity to the HMM does not make the estimation problems any easier. Indeed, the estimation challenges only become exacerbated.

TABLE I

A COMPARISON OF ACCURACY AND ERROR PERCENTAGES OF VARIOUS HMM MODELS, USING THE FULL AND REDUCED LEXICONS.

| Method | full lexicon | | reduced lexicon | |
|---|---|---|---|---|
| | accu. | error | accu. | error |
| Standard | 81.32 | 18.68 | 93.80 | 6.20 |
| Constrained | 90.47 | 9.53 | 94.07 | 5.93 |
| Similarity | 87.44 | 12.56 | 94.06 | 5.94 |
| Combined | 90.03 | 9.97 | 94.68 | 5.32 |
| Weak Const. | 85.93 | 14.07 | 93.97 | 6.03 |

## IV. MARGINALLY CONSTRAINED HMMS

To attempt to improve the quality of the estimates, our first idea is to modify the EM training procedure to improve the quality of the *tag→tag* transition model. Here we consider a very natural constraint: we would like to force the learned HMM to maintain a specified marginal distribution over the tag probabilities. That is, assume we are given a target distribution over the tags, $\pi_t = P(t)$. (Initially we just use the probability of each tag in the complete Treebank; below we show how to estimate these probabilities without tagged data.) Then given this target $\boldsymbol{\pi}$, we would like to modify the EM algorithm to ensure that the raw frequency of each tag $t$ matches $\pi_t$. Although this is an obvious and simple idea, we expect it to improve the quality of the transition model by keeping the transition parameters $\mathbf{a}$ to reasonable values.[2]

The standard EM algorithm can be modified to achieve a specified tag marginal as follows. First, to ensure $\boldsymbol{\pi}$ is indeed the marginal tag distribution, we need to ensure that the transition model, $\mathbf{a}$, maintains $\boldsymbol{\pi}$ as its stationary distribution. To achieve this, the only modification required to EM is to change the M-step for the parameters in $\mathbf{a}$—the E-step and the M-step for the $\mathbf{b}$ parameters remain the same.

Focusing on the transition parameters $\mathbf{a}$, the M-step is usually derived by maximizing the relevant part of the expected complete log likelihood

$$\sum_{j=1}^{N} \sum_{i=1}^{j_n} \sum_{st} \rho(j, i-1, i, s, t) \log a_{st} \tag{1}$$

$$\text{subject to } \sum_{t} a_{st} = 1 \text{ for all } s \tag{2}$$

where the weights

$$\rho(j, i-1, i, s, t) \;=\; P(t_{i-1}^{(j)} = s, t_i^{(j)} = t | \mathbf{w}^{(j)}\mathbf{ab})$$

are the marginal probabilities of tag pair $st$ given the observed sequence $\mathbf{w}^{(j)}$. (These weights are the expected sufficient statistics computed during the E-step using the forward-backward algorithm.) To modify this M-step we simply have to add the constraint that

$$\sum_{s} \pi_s a_{st} = \pi_t \text{ for all } t \tag{3}$$

The objective (1), along with the constraints (2) and (3), specifies a constrained optimization problem over the parameters $\mathbf{a}$. The Lagrangian for this problem is given by

$$\sum_{j=1}^{N} \sum_{i=1}^{j_n} \sum_{st} \rho(j, i-1, i, s, t) \log a_{st}$$
$$- \sum_{s} \lambda_s \left(1 - \sum_{t} a_{st}\right)$$
$$- \sum_{t} \mu_t \left(\pi_t - \sum_{s} \pi_s a_{st}\right)$$

where $\lambda_s$ are the Lagrange multipliers for the constraints $\sum_t a_{st} = 1$, and $\mu_t$ are the Lagrange multipliers for the constraints $\sum_s \pi_s a_{st} = \pi_t$. Taking derivatives and solving for a critical point shows that the transition parameters $\mathbf{a}$ are given by

$$a_{st} \;=\; \frac{\sum_{j=1}^{N} \sum_{i=1}^{j_n} \rho(j, i-1, i, s, t)}{\lambda_s + \mu_t \pi_s} \tag{4}$$

To compute the final solution $\mathbf{a}$ for the M-step, one only needs to find values for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that (2) and (3) are satisfied when the expression (4) is substituted for $a_{st}$ in the constraints. Unfortunately, this problem does not have a closed form solution. Nevertheless, it is easy to solve for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ using numerical methods. Currently, we use an optimization library routine (BFGS) which takes less than a second to solve this problem for each M-step.

### *A. Results*

To determine what benefits, if any, we could achieve by adding the marginalization constraint, we repeated the previous experiment using the constrained version of EM. For the constraint $\boldsymbol{\pi}$ we supplied the list of 45 true tag probabilities obtained from the entire Treebank 3 corpus. (We consider a technique for estimating this distribution without tagged data below.) Figure 1 and Table I show the results (see "Constrained"). Here we see that constrained EM, using the same uniform initialization as before, now reduces the word tagging error rate to 9.5%—half the error rate achieved by standard EM (18.7%). We believe that this establishes the best known unsupervised part-of-speech tagging performance by a factor of two (based on using the full, unedited lexicon). In fact, the resulting error rate has dropped to within a factor of two of the best error rates achieved using an artificially ablated lexicon. (For the record, the constrained HMM achieves an error rate of 5.9% using the ablated lexicon.) However, these positive results depend on having a reasonable tag distribution.

## V. SIMILARITY BASED SMOOTHING

The second idea we investigate attempts to improve the lexical *tag→word* probabilities. As observed in Section III, EM tends to over-fit the word emission parameters, since many of these parameters have few associated observations, and yet the M-step treats each one as an independent optimization variable. Specifically, in EM, there is no intrinsic notion of

[2] We recently discovered that a very similar marginalization constraint was considered by [9]. However, the actual training procedure was not described. Merialdo [9] also reports disappointing results on augmenting a supervised tagger by running EM (and constrained EM) on additional untagged data. In our case, we appear to obtain much more positive results for unsupervised learning.

similarity between the parameters beyond how they affect the training objective.

Although one could imagine trying to constrain the M-step to respect a certain marginal distribution over word frequencies (analogous to Section IV), our intuition instead is that parameter smoothing between similar words should have a more obvious, beneficial effect.

Although there is a significant literature on measuring word similarity in large text copora [16], [17], [18], we simply adopt the approach of [18] which appears to work well in many different scenarios [19], [20]. To measure word similarity, we first construct a feature vector $\mathbf{f}_w$ for each word $w$, which consists of the pointwise mutual information between $w$ and a "context" $c$. Each context corresponds to a non-stop word occurring immediately to the left or right of $w$ (left and right are two different contexts) where stop-words are skipped.[3] Thus, feature vector entries are given by

$$f_w(c) = \log \frac{P(w \text{ and } c)}{P(w)P(c)}$$

for each possible context $c$.[4] To compute these probabilities we used a collection of auxiliary texts (the complete Acquaint, Reuters, and Tipster corpora), totaling 15GB of untagged data. In our experiments, We took the 100,000 most frequent words as features.

Once the feature vectors have been determined, we compute the *similarity* between two words $w_1$ and $w_2$ by taking the cosine of their feature vectors

$$sim(w_1, w_2) = \frac{\mathbf{f}_{w_1} \cdot \mathbf{f}_{w_2}}{||\mathbf{f}_{w_1}||||\mathbf{f}_{w_2}||}$$

This is motivated by the fact that if two words share the same contextual associations with the remainder of the vocabulary (as measured by the pointwise mutual information) they are likely to be similar.

We use this similarity measure to smooth the lexical parameters $b_{tw} = P(w|t)$ by taking similarity weighted averages, analogous to the approach taken in [16], [17]. Specifically, we compute smoothed emission probabilities by

$$\begin{aligned}
\hat{b_{wt}} &= \hat{P}(t|w)P(w)/\hat{P}(t), \text{ where} \\
\hat{P}(t|w) &= \frac{\sum_{w' \in S(w)} sim(w', w) P_{uni}(t|w')}{Z} \\
P_{uni}(t|w') &= 1/\text{\# legal tags for word } w' \text{ in lexicon} \\
\hat{P}(t) &= \sum_w \hat{P}(t|w)P(w), \qquad (5)
\end{aligned}$$

where Z is the normalizer and S(w) is the set of top-50 most similar words of w (for the reported results in this paper, 0.04 is used as a threshold to determine the word pair is similar or not).

### A. Results

To investigate the benefit of smoothing the emission parameters, we repeated the initial HMM experiment using the standard EM algorithm (without the marginalization constraint of Section IV) to isolate the effects of the two modifications. In this case, we initialized the emission parameters $b_{wt}$ to $b_{wt} = \hat{b}_{wt}$, and initialized the tag transition parameters $a_{st}$ to $s_{St} = \hat{P}(t)$ for all $st$.

Figure 1 and Table I show the results of running EM with similarity smoothed initialization (see "Similarity"). Interestingly, before running any EM iterations (just using the initial smoothed parameters) the HMM tagger achieves an error rate of 17.7%, which already surpasses that of the fully trained standard HMM. After a small number of EM iterations, error is reduced to 12.6% (and then begins to rise again as over-fitting begins to occur). Nevertheless, this is a significant (one third) reduction in error over standard EM, which although not as dramatic as the marginal tag constraint, remains significant.

## VI. Combined Results

Given that each of the two proposed improvements appears to work reasonably well, it is natural to consider combining them. However, here we obtain the disappointing result that the combined EM training algorithm, using both the tag marginalization constraint and similarity smoothed emission initialization, does not improve on the previous error rates (Figure 1 and Table I; see "Combined"). In this case, the combined method obtains an error of 9.97%, which is better than the similarity result but slightly worse than the original constrained result. We speculate that the 9.5% error obtained by constrained EM is sufficiently close to the limit of the HMM model that further improvement requires more subtle combinations of the two ideas than those presented here. An immediate research direction is to incorporate similarity smoothing throughout the EM training process, rather than just at initialization.

Nevertheless, we can find an immediately useful application of word similarity to the marginally constrained EM procedure: if the true marginal distribution over tags $P(t)$ is not known *a priori*, it can in fact be estimated using the approximation $\hat{P}(t)$ given in (5). (Note that previously we used the oracle choice based on the true tag frequencies, which may not always be readily available.) Repeating the previous experiment using this weaker estimate of $P(t)$ as the marginal constraint (Figure 1 and Table I; see "Weak Const.") we notice some degradation in performance over using the true marginal—the error rate is now only reduced to 14.1% instead of 9.5%. However, this remains a significant improvement over the baseline 18.7%.

## VII. Conclusion

Contrary to the current trend, we have focused on a very simple prediction model (the standard HMM) and attempted to improve basic estimation quality, rather than propose more complex prediction architectures. Nevertheless, even given the simplicity of this model, the estimation issues are complex and non-obvious. For unsupervised part-of-speech tagging

[3] We used the 100 most frequent words as stop words.

[4] To reduce bias toward infrequent words and contexts we adjust the pointwise mutual information calculation in the same manner as [19].

with HMMs, we presented two improvements, each of which individually obtains state of the art performance on the full unedited lexicon.

Although the error rates we achieve are still higher than those obtained by supervised taggers, the unsupervised error reductions appear to be significant—suggesting that estimation, in addition to model building, could remain an important issue for future research. Of course, given progress on improving estimation quality, one would hope that similar improvements could also be obtained for more complex models. Perhaps through a combination of greater context and improved estimation, the performance of fully automated unsupervised learning could ultimately be made to challenge that of supervised taggers.

## References

[1] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proceedings of EMNLP-1996*, 1996.

[2] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy methods for information extraction and segmentation," in *Proceedings of ICML-2000*, 2000.

[3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML-2001*, 2001.

[4] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Proceedings of EMNLP-2002*, 2002.

[5] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL-2003*, Edmonton, 2003.

[6] M. Banko and R. Moore, "Part-of-speech tagging in context," in *Proceedings of COLING-2004*, Geneva, Switzerland, 2004, pp. 556–561.

[7] N. A. Smith and J. Eisner, "Annealing techniques for unsupervised statistical language learning," in *Proceedings of ACL-2004*, Barcelona, July 2004. [Online]. Available: http://cs.jhu.edu/ jason/papers/

[8] J. Kupiec, "Robust part-of-speech tagging using a hidden markov model," *Computer Speech and Language*, vol. 6, pp. 225–242, 1992.

[9] B. Merialdo, "Tagging english text with a probabilistic model," *Computational Linguistics*, vol. 20, no. 2, pp. 155–171, 1994.

[10] D. Elworthy, "Does baum-welch re-estimation help taggers?" in *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart, Germany, 1994, pp. 53–58.

[11] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992. [Online]. Available: citeseer.ist.psu.edu/cutting92practical.html

[12] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, 1988, pp. 136–143.

[13] E. Brill, "Unsupervised learning of disambiguation rules for part of speech tagging," in *Proceedings of the Third Workshop on Very Large Corpora*, D. Yarovsky and K. Church, Eds. Somerset, New Jersey: Association for Computational Linguistics, 1995, pp. 1–13. [Online]. Available: citeseer.ist.psu.edu/brill95unsupervised.html

[14] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[16] L. Lee, "Measures of distributional similarity," in *Proceedings of ACL-1999*, 1999, pp. 25–32.

[17] I. Dagan, L. Lee, and F. C. N. Pereira, "Similarity-based models of word cooccurrence probabilities," *Machine Learning*, vol. 34, no. 1-3, pp. 43–69, 1999. [Online]. Available: citeseer.ist.psu.edu/dagan99similaritybased.html

[18] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of ICML-1998*. Madison, Wisconsin: Morgan Kaufmann, 1998, pp. 296–304.

[19] ——, "Automatic retrieval and clustering of similar words," in *Proceedings of COLING/ACL-1998*, Montreal, 1998, pp. 768–774.

[20] P. Pantel and D. Lin, "Discovering word senses from text," in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 613–619.