Unsupervised Multilingual Learning for POS Tagging

Benjamin Snyder and **Tahira Naseem** and **Jacob Eisenstein** and **Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge MA 02139
{bsnyder, tahira, jacobe, regina}@csail.mit.edu

Abstract

We demonstrate the effectiveness of multilingual learning for unsupervised part-of-speech tagging. The key hypothesis of multilingual learning is that by combining cues from multiple languages, the structure of each becomes more apparent. We formulate a hierarchical Bayesian model for jointly predicting bilingual streams of part-of-speech tags. The model learns language-specific features while capturing cross-lingual patterns in tag distribution for aligned words. Once the parameters of our model have been learned on bilingual parallel data, we evaluate its performance on a held-out monolingual test set. Our evaluation on six pairs of languages shows consistent and significant performance gains over a state-of-the-art monolingual baseline. For one language pair, we observe a relative reduction in error of 53%.

1 Introduction

In this paper, we explore the application of multilingual learning to part-of-speech tagging when no annotation is available. This core task has been studied in an unsupervised monolingual framework for over a decade and is still an active area of research. In this paper, we demonstrate the effectiveness of multilingual learning when applied to both closely related and distantly related language pairs. We further analyze the language features which lead to robust bilingual performance.

The fundamental idea upon which our work is based is that the patterns of ambiguity inherent in part-of-speech tag assignments differ across languages. At the lexical level, a word with part-of-speech tag ambiguity in one language may correspond to an unambiguous word in the other language. For example, the word "can" in English may function as an auxiliary verb, a noun, or a regular verb. However, each of the corresponding functions in Serbian is expressed with a distinct lexical item. Languages also differ in their patterns of structural ambiguity. For example, the presence of an article in English greatly reduces the ambiguity of the succeeding tag. In Serbian, a language without articles, this constraint is obviously absent. The key idea of multilingual learning is that by combining cues from multiple languages, the structure of each becomes more apparent.

While multilingual learning can address ambiguities in each language, it must be flexible enough to accommodate cross-lingual variations such as tag inventory and syntactic structure. As a result of such variations, two languages often select and order their tags differently even when expressing the same meaning. A key challenge of multilingual learning is to model language-specific structure while allowing information to flow between languages.

We jointly model bilingual part-of-speech tag sequences in a hierarchical Bayesian framework. For each word, we posit a hidden tag state which generates the word as well as the succeeding tag. In addition, the tags of words with common semantic or syntactic function in parallel sentences are combined into bilingual nodes representing the tag pair. These joined nodes serve as anchors that create probabilistic dependencies between the tag se-

quences in each language. We use standard tools from machine translation to discover aligned word-pairs, and thereafter our model treats the alignments as observed data.

Our model structure allows language-specific tag inventories. Additionally, it assumes only that the tags at joined nodes are *correlated*; they need not be identical. We factor the conditional probabilities of joined nodes into two individual transition probabilities as well as a coupling probability. We define priors over the transition, emission, and coupling parameters and perform Bayesian inference using Gibbs sampling and the Metropolis-Hastings algorithm.

We evaluate our model on a parallel corpus of four languages: English, Bulgarian, Serbian, and Slovene. For each of the six language pairs, we train a bilingual model on this corpus, and evaluate it on held-out monolingual test sets. Our results show consistent improvement over a monolingual baseline for all languages and all pairings. In fact, for one language pair – Serbian and Slovene – the error is reduced by over 53%. Moreover, the multilingual model significantly reduces the gap between unsupervised and supervised performance. For instance, in the case of Slovene this gap is reduced by 71%. We also observe significant variation in the level of improvement across language pairs. We show that a cross-lingual entropy measure corresponds with the observed differentials in performance.

2 Related Work

Multilingual Learning A number of approaches for multilingual learning have focused on inducing cross-lingual structures, with applications to machine translation. Examples of such efforts include work on the induction of synchronous grammars (Wu and Wong, 1998; Chiang, 2005) and learning multilingual lexical resources (Genzel, 2005).

Another thread of work using cross-lingual links has been in word-sense disambiguation, where senses of words can be *defined* based on their translations (Brown et al., 1991; Dagan et al., 1991; Resnik and Yarowsky, 1997; Ng et al., 2003).

When annotations for a task of interest are available in a source language but are missing in the target language, the annotations can be projected across a parallel corpus (Yarowsky et al., 2000; Diab and Resnik, 2002; Padó and Lapata, 2006; Xi and Hwa, 2005). In fact, projection methods have been used to train highly accurate part-of-speech taggers (Yarowsky and Ngai, 2001; Feldman et al., 2006). In contrast, our own work assumes that annotations exist for neither language.

Finally, there has been recent work on applying unsupervised multilingual learning to morphological segmentation (Snyder and Barzilay, 2008). In this paper, we demonstrate that unsupervised multilingual learning can be successfully applied to the sentence-level task of part-of-speech tagging.

Unsupervised Part-of-Speech Tagging Since the work of Merialdo (1994), the HMM has been the model of choice for unsupervised tagging (Banko and Moore, 2004). Recent advances in these approaches include the use of a fully Bayesian HMM (Johnson, 2007; Goldwater and Griffiths, 2007). In very recent work, Toutanova and Johnson (2008) depart from this framework and propose an LDA-based generative model that groups words through a latent layer of ambiguity classes thereby leveraging morphological features. In addition, a number of approaches have focused on developing discriminative approaches for unsupervised and semi-supervised tagging (Smith and Eisner, 2005; Haghighi and Klein, 2006).

Our focus is on developing a simple model that effectively incorporates multilingual evidence. We view this direction as orthogonal to refining monolingual tagging models for any particular language.

3 Model

We propose a bilingual model for unsupervised part-of-speech tagging that jointly tags parallel streams of text in two languages. Once the parameters have been learned using an untagged bilingual parallel text, the model is applied to a held-out monolingual test set.

Our key hypothesis is that the patterns of ambiguity found in each language at the part-of-speech level will differ in systematic ways; by considering multiple language simultaneously, the total inherent ambiguity can be reduced in each language. The model is designed to permit information to flow across the

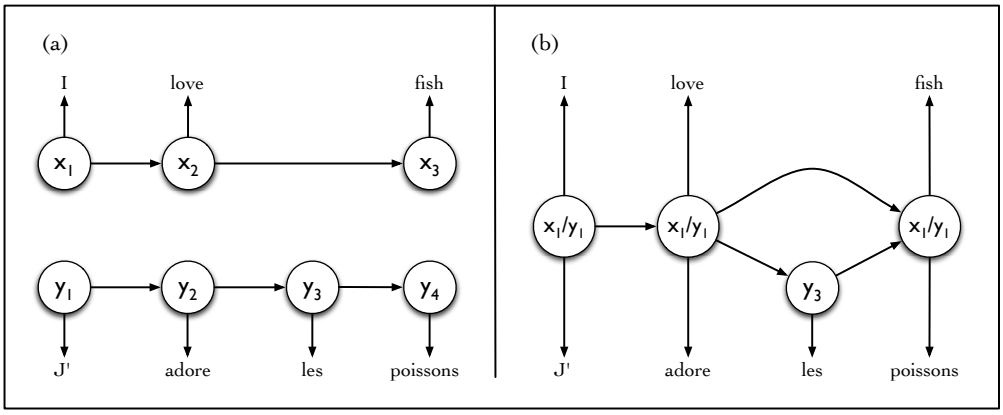

Figure 1: (a) Graphical structure of two standard monolingual HMM's. (b) Graphical structure of our bilingual model based on word alignments.

language barrier, while respecting language-specific idiosyncrasies such as tag inventory, selection, and order. We assume that for pairs of words that share similar semantic or syntactic function, the associated tags will be statistically correlated, though not necessarily identical. We use such word pairs as the bilingual anchors of our model, allowing cross-lingual information to be shared via joint tagging decisions. We use standard tools from machine translation to identify these aligned words, and thereafter our model treats them as fixed and observed data. To avoid cycles, we remove crossing edges from the alignments.

For unaligned parts of the sentence, the tag and word selections are identical to standard monolingual HMM's. Figure 1 shows an example of the bilingual graphical structure we use, in comparison to two independent monolingual HMM's.

We formulate a hierarchical Bayesian model that exploits both language-specific and cross-lingual patterns to explain the observed bilingual sentences. We present a generative story in which the observed words are produced by the hidden tags and model parameters. In Section 4, we describe how to infer the posterior distribution over these hidden variables, given the observations.

3.1 Generative Model

Our generative model assumes the existence of two tagsets, T and T', and two vocabularies W and W', one of each for each language. For ease of exposition, we formulate our model with bigram tag dependencies. However, in our experiments we used a trigram model, which is a trivial extension of the model discussed here and in the next section.

1. For each tag $t \in T$, draw a *transition* distribution ϕ_t over tags T, and an *emission* distribution θ_t over words W, both from symmetric Dirichlet priors.[1]

2. For each tag $t \in T'$, draw a *transition* distribution ϕ'_t over tags T', and an *emission* distribution θ'_t over words W', both from symmetric Dirichlet priors.

3. Draw a bilingual *coupling* distribution ω over tag pairs $T \times T'$ from a symmetric Dirichlet prior.

4. For each bilingual parallel sentence:

   (a) Draw an alignment a from an alignment distribution A (see the following paragraph for formal definitions of a and A),

   (b) Draw a bilingual sequence of part-of-speech tags $(x_1, ..., x_m),\ (y_1, ..., y_n)$ according to:
   $P(x_1, ..., x_m,\ \ y_1, ..., y_n | a, \phi, \phi', \omega)$. [2]
   This joint distribution is given in equation 1.

[1] The Dirichlet is a probability distribution over the simplex, and is conjugate to the multinomial (Gelman et al., 2004).

[2] Note that we use a special end state rather than explicitly modeling sentence length. Thus the values of m and n depend on the draw.

(c) For each part-of-speech tag x_i in the first language, emit a word from W: $e_i \sim \theta_{x_i}$,

(d) For each part-of-speech tag y_j in the second language, emit a word from W': $f_j \sim \theta'_{y_j}$.

We define an alignment a to be a set of one-to-one integer pairs with no crossing edges. Intuitively, each pair $(i, j) \in a$ indicates that the words e_i and f_j share some common role in the bilingual parallel sentences. In our experiments, we assume that alignments are directly observed and we hold them fixed. From the perspective of our generative model, we treat alignments as drawn from a distribution A, about which we remain largely agnostic. We only require that A assign zero probability to alignments which either: *(*i) align a single index in one language to multiple indices in the other language or *(*ii) contain crossing edges. The resulting alignments are thus one-to-one, contain no crossing edges, and may be sparse or even possibly empty. Our technique for obtaining alignments that display these properties is described in Section 5.

Given an alignment a and sets of transition parameters ϕ and ϕ', we factor the conditional probability of a bilingual tag sequence $(x_1, ...x_m)$, $(y_1, ..., y_n)$ into transition probabilities for unaligned tags, and joint probabilities over aligned tag pairs:

$$P(x_1, ..., x_m,\ y_1, ..., y_n | a, \phi, \phi', \omega) = \prod_{\text{unaligned } i} \phi_{x_{i-1}}(x_i) \cdot \prod_{\text{unaligned } j} \phi'_{y_{j-1}}(y_j) \cdot \prod_{(i,j)\in a} P(x_i, y_j | x_{i-1}, y_{j-1}, \phi, \phi', \omega) \quad (1)$$

Because the alignment contains no crossing edges, we can model the tags as generated sequentially by a stochastic process. We define the distribution over aligned tag pairs to be a product of each language's transition probability and the coupling probability:

$$P(x_i, y_j | x_{i-1}, y_{j-1}, \phi, \phi', \omega) = \frac{\phi_{x_{i-1}}(x_i)\ \phi'_{y_{j-1}}(y_j)\ \omega(x_i, y_j)}{Z} \quad (2)$$

The normalization constant here is defined as:

$$Z = \sum_{x,y} \phi_{x_{i-1}}(x)\ \phi'_{y_{j-1}}(y)\ \omega(x, y)$$

This factorization allows the language-specific transition probabilities to be shared across aligned and unaligned tags. In the latter case, the addition of the coupling parameter ω gives the tag pair an additional role: that of multilingual anchor. In essence, the probability of the aligned tag pair is a product of three experts: the two transition parameters and the coupling parameter. Thus, the combination of a high probability transition in one language and a high probability coupling can resolve cases of inherent transition uncertainty in the other language. In addition, any one of the three parameters can "veto" a tag pair to which it assigns low probability.

To perform inference in this model, we predict the bilingual tag sequences with maximal probability given the observed words and alignments, while integrating over the transition, emission, and coupling parameters. To do so, we use a combination of sampling-based techniques.

4 Inference

The core element of our inference procedure is Gibbs sampling (Geman and Geman, 1984). Gibbs sampling begins by randomly initializing all unobserved random variables; at each iteration, each random variable z_i is sampled from the conditional distribution $P(z_i|\mathbf{z}_{-i})$, where $\mathbf{z}_{-i}$ refers to all variables other than z_i. Eventually, the distribution over samples drawn from this process will converge to the unconditional joint distribution $P(\mathbf{z})$ of the unobserved variables. When possible, we avoid explicitly sampling variables which are not of direct interest, but rather integrate over them—this technique is known as "collapsed sampling," and can reduce variance (Liu, 1994).

We sample: *(i)* the bilingual tag sequences $(\mathbf{x}, \mathbf{y})$, *(ii)* the two sets of transition parameters ϕ and ϕ', and *(iii)* the coupling parameter ω. We integrate over the emission parameters θ and θ', whose priors are Dirichlet distributions with hyperparameters θ_0 and θ'_0. The resulting emission distribution over words e_i, given the other words $\mathbf{e}_{-i}$, the tag sequences $\mathbf{x}$

and the emission prior θ_0, can easily be derived as:

$$P(e_i|\mathbf{x}, \mathbf{e}_{-i}, \theta_0) = \int_{\theta_{x_i}} \theta_{x_i}(e_i)\, P(\theta_{x_i}|\theta_0)\, d\theta_{x_i} = \frac{n(x_i, e_i) + \theta_0}{n(x_i) + W_{x_i}\theta_0} \quad (3)$$

Here, $n(x_i)$ is the number of occurrences of the tag x_i in $\mathbf{x}_{-i}$, $n(x_i, e_i)$ is the number of occurrences of the tag-word pair (x_i, e_i) in $(\mathbf{x}_{-i}, \mathbf{e}_{-i})$, and W_{x_i} is the number of word types in the vocabulary W that can take tag x_i. The integral is tractable due to Dirichlet-multinomial conjugacy (Gelman et al., 2004).

We will now discuss, in turn, each of the variables that we sample. Note that in all cases we condition on the other sampled variables as well as the observed words and alignments, $\mathbf{e}$, $\mathbf{f}$ and a, which are kept fixed throughout.

4.1 Sampling Part-of-speech Tags

This section presents the conditional distributions that we sample from to obtain the part-of-speech tags. Depending on the alignment, there are several scenarios. In the simplest case, both the tag to be sampled and its succeeding tag are not aligned to any tag in the other language. If so, the sampling distribution is identical to the monolingual case, including only terms for the emission (defined in equation 3), and the preceding and succeeding transitions:

$$P(x_i|\mathbf{x}_{-i}, \mathbf{y}, \mathbf{e}, \mathbf{f}, a, \phi, \phi', \omega, \theta_0, \theta_0') \propto P(e_i|\mathbf{x}, \mathbf{e}_{-i}, \theta_0)\ \phi_{x_{i-1}}(x_i)\ \phi_{x_i}(x_{i+1}).$$

For an aligned tag pair (x_i, y_j), we sample the identity of the tags jointly. By applying the chain rule we obtain terms for the emissions in both languages and a joint term for the transition probabilities:

$$P(x_i, y_j|\mathbf{x}_{-i}, \mathbf{y}_{-j}, \mathbf{e}, \mathbf{f}, a, \phi, \phi', \omega, \theta_0, \theta_0') \propto P(e_i|\mathbf{x}, \mathbf{e}_{-i}, \theta_0) P(f_j|\mathbf{y}, \mathbf{f}_{-j}, \theta_0') P(x_i, y_j|\mathbf{x}_{-i}, \mathbf{y}_{-j}, a, \phi, \phi', \omega)$$

The expansion of the joint term depends on the alignment of the succeeding tags. In the case that the successors are not aligned, we have a product of the bilingual coupling probability and four transition probabilities (preceding and succeeding transitions in each language):

$$P(x_i, y_j|\mathbf{x}_{-i}, \mathbf{y}_{-j}, a, \phi, \phi', \omega) \propto \omega(x_i, y_j)\phi_{x_{i-1}}(x_i)\ \phi'_{y_{j-1}}(y_j)\ \phi_{x_i}(x_{i+1})\ \phi'_{y_j}(y_{j+1})$$

Whenever one or more of the succeeding tags is aligned, the sampling formulas must account for the effect of the sampled tag on the joint probability of the succeeding tags, which is no longer a simple multinomial transition probability. We give the formula for one such case—when we are sampling an aligned tag pair (x_i, y_j), whose succeeding tags (x_{i+1}, y_{j+1}) are also aligned to one another:

$$P(x_i, y_j|\mathbf{x}_{-i}, \mathbf{y}_{-j}, a, \phi, \phi', \omega) \propto \omega(x_i, y_j) \cdot \phi_{x_{i-1}}(x_i)\, \phi'_{y_{j-1}}(y_j) \left[\frac{\phi_{x_i}(x_{i+1})\, \phi'_{y_j}(y_{j+1})}{\sum_{x,y} \phi_{x_i}(x)\, \phi'_{y_j}(y)\, \omega(x, y)}\right]$$

Similar equations can be derived for cases where the succeeding tags are not aligned to each other, but to other tags.

4.2 Sampling Transition Parameters and the Coupling Parameter

When computing the joint probability of an aligned tag pair (Equation 2), we employ the transition parameters ϕ, ϕ' and the coupling parameter ω in a normalized product. Because of this, we can no longer regard these parameters as simple multinomials, and thus can no longer sample them using the standard closed formulas.

Instead, to resample these parameters, we resort to the Metropolis-Hastings algorithm as a subroutine within Gibbs sampling (Hastings, 1970). Metropolis-Hastings is a Markov chain sampling technique that can be used when it is impossible to directly sample from the posterior. Instead, samples are drawn from a *proposal* distribution and then stochastically accepted or rejected on the basis of: their likelihood, their probability under the proposal distribution, and the likelihood and proposal probability of the previous sample.

We use a form of Metropolis-Hastings known as an *independent sampler*. In this setup, the proposal distribution does not depend on the value of the previous sample, although the accept/reject decision

does depend on the previous model likelihood. More formally, if we denote the proposal distribution as $Q(z)$, the target distribution as $P(z)$, and the previous sample as z, then the probability of accepting a new sample $z^* \sim Q$ is set at:

$$\min\left\{1, \frac{P(z^*)\,Q(z)}{P(z)\,Q(z^*)}\right\}$$

Theoretically any non-degenerate proposal distribution may be used. However, a higher acceptance rate and faster convergence is achieved when the proposal Q is a close approximation of P. For a particular transition parameter ϕ_x, we define our proposal distribution Q to be Dirichlet with parameters set to the bigram counts of the tags following x in the sampled tag data. Thus, the proposal distribution for ϕ_x has a mean proportional to these counts, and is thus likely to be a good approximation to the target distribution.

Likewise for the coupling parameter ω, we define a Dirichlet proposal distribution. This Dirichlet is parameterized by the counts of aligned tag pairs (x, y) in the current set of tag samples. Since this sets the mean of the proposal to be proportional to these counts, this too is likely to be a good approximation to the target distribution.

4.3 Hyperparameter Re-estimation

After every iteration of Gibbs sampling the hyperparameters θ_0 and θ_0' are re-estimated using a single Metropolis-Hastings move. The proposal distribution is set to a Gaussian with mean at the current value and variance equal to one tenth of the mean.

5 Experimental Set-Up

Our evaluation framework follows the standard procedures established for unsupervised part-of-speech tagging. Given a tag dictionary (i.e., a set of possible tags for each word type), the model has to select the appropriate tag for each token occurring in a text. We also evaluate tagger performance when only incomplete dictionaries are available (Smith and Eisner, 2005; Goldwater and Griffiths, 2007). In both scenarios, the model is trained only using untagged text.

In this section, we first describe the parallel data and part-of-speech annotations used for system evaluation. Next we describe a monolingual baseline and our procedures for initialization and hyperparameter setting.

Data As a source of parallel data, we use Orwell's novel "Nineteen Eighty Four" in the original English as well as translations to three Slavic languages — Bulgarian, Serbian and Slovene. This data is distributed as part of the Multext-East corpus which is publicly available. The corpus provides detailed morphological annotation at the world level, including part-of-speech tags. In addition a lexicon for each language is provided.

We obtain six parallel corpora by considering all pairings of the four languages. We compute word level alignments for each language pair using Giza++. To generate one-to-one alignments at the word level, we intersect the one-to-many alignments going in each direction and automatically remove crossing edges in the order in which they appear left to right. This process results in alignment of about half the tokens in each bilingual parallel corpus. We treat the alignments as fixed and observed variables throughout the training procedure.

The corpus consists of 94,725 English words (see Table 2). For every language, a random three quarters of the data are used for learning the model while the remaining quarter is used for testing. In the test set, only monolingual information is made available to the model, in order to simulate future performance on non-parallel data.

	Tokens	Tags/Token
SR	89,051	1.41
SL	91,724	1.40
BG	80,757	1.34
EN	94,725	2.58

Table 2: Corpus statistics: SR=Serbian, SL=Slovene, EN=English, BG=Bulgarian

Tagset The Multext-East corpus is manually annotated with detailed morphosyntactic information. In our experiments, we focus on the main syntactic category encoded as a first letter of the labels. The annotation distinguishes between 13 parts-of-speech, of which 11 are common for all languages

	Random	Monolingual Unsupervised	Monolingual Supervised	Trigram Entropy
EN	56.24	90.71	96.97	1.558
BG	82.68	88.88	96.96	1.708
SL	84.70	87.41	97.31	1.703
SR	83.41	85.05	96.72	1.789

Table 1: Monolingual tagging accuracy for English, Bulgarian, Slovene, and Serbian for two unsupervised baselines (random tag selection and a Bayesian HMM (Goldwater and Griffiths, 2007)) as well as a supervised HMM. In addition, the trigram part-of-speech tag entropy is given for each language.

in our experiments.[3]

In the Multext-East corpus, punctuation marks are not annotated. We expand the tag repository by defining a separate tag for all punctuation marks. This allows the model to make use of any transition or coupling patterns involving punctuation marks. We do not consider punctuation tokens when computing model accuracy.

Table 2 shows the tag/token ratio for these languages. For Slavic languages, we use the tag dictionaries provided with the corpus. For English, we use a different process for dictionary construction. Using the original dictionary would result in the tag/token ratio of 1.5, in comparison to the ratio of 2.3 observed in the Wall Street Journal (WSJ) corpus. To make our results on English tagging more comparable to previous benchmarks, we expand the original dictionary of English tags by merging it with the tags from the WSJ dictionary. This process results in a tag/token ratio of 2.58, yielding a slightly more ambiguous dictionary than the one used in previous tagging work. [4]

Monolingual Baseline As our monolingual baseline we use the unsupervised Bayesian HMM model of Goldwater and Griffiths (2007) (BHMM1). This model modifies the standard HMM by adding priors and by performing Bayesian inference. Its is in line with state-of-the-art unsupervised models. This model is a particulary informative baseline, since our model reduces to this baseline model when there are no alignments in the data. This implies that any performance gain over the baseline can only be attributed to the multilingual aspect of our model. We used our own implementation after verifying that its performance on WSJ was identical to that reported in (Goldwater and Griffiths, 2007).

Supervised Performance In order to provide a point of comparison, we also provide supervised results when an annotated corpus is provided. We use the standard supervised HMM with Viterbi decoding.

Training and Testing Framework Initially, all words are assigned tags randomly from their tag dictionaries. During each iteration of the sampler, aligned tag pairs and unaligned tags are sampled from their respective distributions given in Section 4.1 above. The hyperparameters θ_0 and θ'_0 are initialized with the values learned during monolingual training. They are re-estimated after every iteration of the sampler using the Metropolis Hastings algorithm. The parameters ϕ and ϕ' are initially set to trigram counts and the ω parameter is set to tag pair counts of aligned pairs. After every 40 iterations of the sampler, a Metropolis Hastings subroutine is invoked that re-estimates these parameters based on the current counts. Overall, the algorithm is run for 1000 iterations of tag sampling, by which time the resulting log-likelihood converges to stable values. Each Metropolis Hastings subroutine samples 20 values, with an acceptance ratio of around 1/6, in line with the standard recommended values.

After training, trigram and word emission probabilities are computed based on the counts of tags assigned in the final iteration. For smoothing, the final sampled values of the hyperparameters are used. The highest probability tag sequences for each monolingual test set are then predicted using trigram Viterbi decoding. We report results averaged over five complete runs of all experiments.

[3]The remaining two tags are Particle and Determiner; The English tagset does not include *Particle* while the other three languages Serbian, Slovene and Bulgarian do not have *Determiner* in their tagset.

[4]We couldn't perform the same dictionary expansion for the Slavic languages due to a lack of additional annotated resources.

6 Results

Complete Tag Dictionary In our first experiment, we assume that a complete dictionary listing the possible tags for every word is provided in each language. Table 1 shows the monolingual results of a random baseline, an unsupervised Bayesian HMM and a supervised HMM. Table 3 show the results of our bilingual models for different language pairings while repeating the monolingual unsupervised results from Table 1 for easy comparison. The final column indicates the absolute gain in performance over this monolingual baseline.

Across all language pairs, the bilingual model consistently outperforms the monolingual baseline. All the improvements are statistically significant by a Fisher sign test at $p < 0.05$. For some language pairs, the gains are quite high. For instance, the pairing of Serbian and Slovene (two closely related languages) yields absolute improvements of 6.7 and 7.7 percentage points, corresponding to relative reductions in error of 51.4% and 53.2%. Pairing Bulgarian and English (two distantly related languages) also yields large gains: 5.6 and 1.3 percentage points, corresponding to relative reductions in error of 50% and 14%, respectively.[5]

When we compare the best bilingual result for each language (Table 3, in bold) to the monolingual supervised results (Table 1), we find that for all languages the gap between supervised and unsupervised learning is reduced significantly. For English, this gap is reduced by 21%. For the Slavic languages, the supervised-unsupervised gap is reduced by even larger amounts: 57%, 69%, and 78% for Serbian, Bulgarian, and Slovene respectively.

While all the languages benefit from the bilingual learning framework, some language combinations are more effective than others. Slovene, for instance, achieves a large improvement when paired with Serbian ($+7.7$), a closely related Slavic language, but only a minor improvement when coupled with English ($+1.3$). On the other hand, for Bulgarian, the best performance is achieved when coupling with English ($+5.6$) rather than with closely related Slavic languages ($+3.1$ and $+2.4$). As these results show, an optimal pairing cannot be predicted based solely on the family connection of paired languages.

	Entropy	Monolingual	Bilingual	Absolute Gain
EN	0.566	90.71	91.01	+0.30
SR	0.554	85.05	90.06	+5.03
EN	0.578	90.71	92.00	+1.29
BG	0.543	88.88	**94.48**	**+5.61**
EN	0.571	90.71	**92.01**	**+1.30**
SL	0.568	87.41	88.54	+1.13
SL	0.494	87.41	**95.10**	**+7.69**
SR	0.478	85.05	**91.75**	**+6.70**
BG	0.568	88.88	91.95	+3.08
SR	0.588	85.05	86.58	+1.53
BG	0.579	88.88	90.91	+2.04
SL	0.609	87.41	88.20	+0.79

Table 3: The tagging accuracy of our bilingual models on different language pairs, when a full tag dictionary is provided. The Monolingual Unsupervised results from Table 1 are repeated for easy comparison. The first column shows the cross-lingual entropy of a tag when the tag of the aligned word in the other language is known. The final column shows the absolute improvement over the monolingual Bayesian HMM. The best result for each language is shown in boldface.

To gain a better understanding of this variation in performance, we measured the internal tag entropy of each language as well as the cross-lingual tag entropy of language pairs. For the first measure, we computed the conditional entropy of a tag decision given the previous two tags. Intuitively, this should correspond to the inherent structural uncertainty of part-of-speech decisions in a language. In fact, as Table 1 shows, the trigram entropy is a good indicator of the relative performance of the monolingual baseline. To measure the cross-lingual tag entropies of language pairs, we considered all bilingual aligned tag pairs, and computed the conditional entropy of the tags in one language given the tags in the other language. This measure should indicate the amount of information that one language in a pair can provide the other. The results of this anal-

[5]The accuracy of the monolingual English tagger is relatively high compared to the 87% reported by (Goldwater and Griffiths, 2007) on the WSJ corpus. We attribute this discrepancy to the slight differences in tag inventory used in our dataset. For example, when *Particles* and *Prepositions* are merged in the WSJ corpus (as they happen to be in our tag inventory and corpus), the performance of Goldwater's model on WSJ is similar to what we report here.

	Mono-lingual	Bilingual	Absolute Gain
EN	63.57	68.22	+4.66
SR	41.14	54.73	+13.59
EN	63.57	**71.34**	**+7.78**
BG	53.19	**62.55**	**+9.37**
EN	63.57	66.48	+2.91
SL	49.90	53.77	+3.88
SL	49.90	**59.68**	**+9.78**
SR	41.14	54.08	+12.94
BG	53.19	54.22	+1.04
SR	41.14	**56.91**	**+15.77**
BG	53.19	55.88	+2.70
SL	49.90	58.50	+8.60

Table 4: Tagging accuracy for Bilingual models with reduced dictionary: Lexicon entries are available for only the 100 most frequent words, while all other words become fully ambiguous. The improvement over the monolingual Bayesian HMM trained under similar circumstances is shown. The best result for each language is shown in boldface.

ysis are given in the first column of Table 3. We observe that the cross-lingual entropy is lowest for the Serbian and Slovene pair, corresponding with their large gain in performance. Bulgarian, on the other hand, has lowest cross-lingual entropy when paired with English. This corresponds with the fact that English provides Bulgarian with its largest performance gain. In general, we find that the largest performance gain for any language is achieved when minimizing its cross-lingual entropy.

Reduced Tag Dictionary We also conducted experiments to investigate the impact of the dictionary size on the performance of the bilingual model. Here, we provide results for the realistic scenario where only a very small dictionary is present. Table 4 shows the performance when a tag dictionary for the 100 most frequent words is present in each language. The bilingual model's results are consistently and significantly better than the monolingual baseline for all language pairs.

7 Conclusion

We have demonstrated the effectiveness of multilingual learning for unsupervised part-of-speech tagging. The key hypothesis of multilingual learning is that by combining cues from multiple languages, the structure of each becomes more apparent. We formulated a hierarchical Bayesian model for jointly predicting bilingual streams of tags. The model learns language-specific features while capturing cross-lingual patterns in tag distribution. Our evaluation shows significant performance gains over a state-of-the-art monolingual baseline.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (CAREER grant IIS-0448168 and grant IIS-0835445) and the Microsoft Research Faculty Fellowship. Thanks to Michael Collins, Amir Globerson, Lillian Lee, Yoong Keok Lee, Maria Polinsky and the anonymous reviewers for helpful comments and suggestions. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the NSF.

References

Michele Banko and Robert C. Moore. 2004. Part-of-speech tagging in context. In *Proceedings of the COLING*, pages 556–561.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the ACL*, pages 264–270.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL*, pages 263–270.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the ACL*, pages 130–137.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the ACL*, pages 255–262.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*, pages 549–554.

Andrew Gelman, John B. Carlin, Hal .S. Stern, and Donald .B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

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.

Dmitriy Genzel. 2005. Inducing a multilingual dictionary from a parallel multitext in related languages. In *Proceedings of HLT/EMNLP*, pages 875–882.

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

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. *Proceedings of HLT-NAACL*, pages 320–327.

W. K. Hastings. 1970. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

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

Jun S. Liu. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

Bernard Merialdo. 1994. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *Proceedings of the ACL*, pages 455–462.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL*, pages 1161 – 1168.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86.

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

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the ACL/HLT*, pages 737–745.

Kristina Toutanova and Mark Johnson. 2008. A Bayesian lda-based model for semi-supervised part-of-speech tagging. In *Advances in Neural Information Processing Systems 20*, pages 1521–1528. MIT Press.

Dekai Wu and Hongsing Wong. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of the ACL/COLING*, pages 1408–1415.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of EMNLP*, pages 851 – 858.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the NAACL*, pages 1–8.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2000. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, pages 161–168.