Adding More Languages Improves Unsupervised Multilingual Part-of-Speech Tagging: A Bayesian Non-Parametric Approach

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

Abstract

We investigate the problem of unsupervised part-of-speech tagging when raw parallel data is available in a large number of languages. Patterns of ambiguity vary greatly across languages and therefore even unannotated multilingual data can serve as a learning signal. We propose a non-parametric Bayesian model that connects related tagging decisions across languages through the use of multilingual latent variables. Our experiments show that performance improves steadily as the number of languages increases.

1 Introduction

In this paper we investigate the problem of unsupervised part-of-speech tagging when unannotated parallel data is available in a large number of languages. Our goal is to develop a fully joint multilingual model that scales well and shows improved performance for individual languages as the total *number* of languages increases.

Languages exhibit ambiguity at multiple levels, making unsupervised induction of their underlying structure a difficult task. However, sources of linguistic ambiguity vary across languages. For example, the word *fish* in English can be used as either a verb or a noun. In French, however, the noun *poisson* (fish) is entirely distinct from the verbal form *pêcher* (to fish). Previous work has leveraged this idea by building models for unsupervised learning from aligned bilingual data (Snyder et al., 2008). However, aligned data is often available for *many* languages. The benefits of bilingual learning vary markedly depending on which pair of languages is selected, and without labeled data it is unclear how to determine which supplementary language is most helpful. In this paper, we show that it is possible to leverage all aligned languages simultaneously, achieving accuracy that in most cases outperforms even optimally chosen bilingual pairings.

Even in expressing the same meaning, languages take different syntactic routes, leading to variation in part-of-speech sequences. Therefore, an effective multilingual model must accurately model common linguistic structure, yet remain flexible to the idiosyncrasies of each language. This tension only becomes stronger as additional languages are added to the mix. From a computational standpoint, the main challenge is to ensure that the model scales well as the number of languages increases. Care must be taken to avoid an exponential increase in the parameter space as well as the time complexity of inference procedure.

We propose a non-parametric Bayesian model for joint multilingual tagging. The topology of our model connects tagging decisions within a language as well as across languages. The model scales linearly with the number of languages, allowing us to incorporate as many as are available. For each language, the model contains an HMM-like substructure and connects these substructures to one another by means of cross-lingual latent variables. These variables, which we refer to as *superlingual tags*, capture repeated multilingual patterns and thus reduce the overall uncertainty in tagging decisions.

We evaluate our model on a parallel corpus of eight languages. The model is trained once using all

languages, and its performance is tested separately for each on a held-out monolingual test set. When a complete tag lexicon is provided, our unsupervised model achieves an average accuracy of 95%, in comparison to 91% for an unsupervised monolingual Bayesian HMM and 97.4% for its supervised counterpart. Thus, on average, the gap between unsupervised and supervised monolingual performance is cut by nearly two thirds. We also examined scenarios where the tag lexicon is reduced in size. In all cases, the multilingual model yielded substantial performance gains. Finally, we examined the performance of our model when trained on all possible subsets of the eight languages. We found that performance improves steadily as the number of available languages increases.

2 Related Work

Bilingual Part-of-Speech Tagging Early work on multilingual tagging focused on projecting annotations from an annotated source language to a target language (Yarowsky and Ngai, 2001; Feldman et al., 2006). In contrast, we assume no labeled data at all; our unsupervised model instead symmetrically improves performance for all languages by learning cross-lingual patterns in raw parallel data. An additional distinction is that projection-based work utilizes pairs of languages, while our approach allows for continuous improvement as languages are added to the mix.

In recent work, Snyder et al. (2008) presented a model for unsupervised part-of-speech tagging trained from a bilingual parallel corpus. This bilingual model and the model presented here share a number of similarities: both are Bayesian graphical models building upon hidden Markov models. However, the bilingual model explicitly joins each aligned word-pair into a single coupled state. Thus, the state-space of these joined nodes grows exponentially in the number of languages. In addition, crossing alignments must be removed so that the resulting graph structure remains acyclic. In contrast, our multilingual model posits latent cross-lingual tags without explicitly joining or directly connecting the part-of-speech tags across languages. Besides permitting crossing alignments, this structure allows the model to scale gracefully with the number of languages.

Beyond Bilingual Learning While most work on multilingual learning focuses on bilingual analysis, some models operate on more than one pair of languages. For instance, Genzel (2005) describes a method for inducing a multilingual lexicon from a group of related languages. His model first induces bilingual models for each pair of languages and then combines them. Our work takes a different approach by simultaneously learning from all languages, rather than combining bilingual results.

A related thread of research is multi-source machine translation (Och and Ney, 2001; Utiyama and Isahara, 2006; Cohn and Lapata, 2007) where the goal is to translate from multiple source languages to a single target language. Rather than jointly training all the languages together, these models train bilingual models separately, and then use their output to select a final translation. The selection criterion can be learned at training time since these models have access to the correct translation. In unsupervised settings, however, we do not have a principled means for selecting among outputs of different bilingual models. By developing a joint multilingual model we can automatically achieve performance that rivals that of the best bilingual pairings.

3 Model

We propose a non-parametric directed Bayesian graphical model for multilingual part-of-speech tagging using a parallel corpus. We perform a joint training pass over the corpus, and then apply the parameters learned for each language to a held-out monolingual test set.

The core idea of our model is that patterns of ambiguity vary across languages and therefore even unannotated multilingual data can serve as a learning signal. Our model is able to simultaneously harness this signal from *all* languages present in the corpus. This goal is achieved by designing a single graphical model that connects tagging decisions within a language as well as across languages.

The model contains language-specific HMM substructures connected to one another by cross-lingual latent variables spanning two or more languages. These variables, which we refer to as *superlingual tags*, capture repeated cross-lingual patterns and

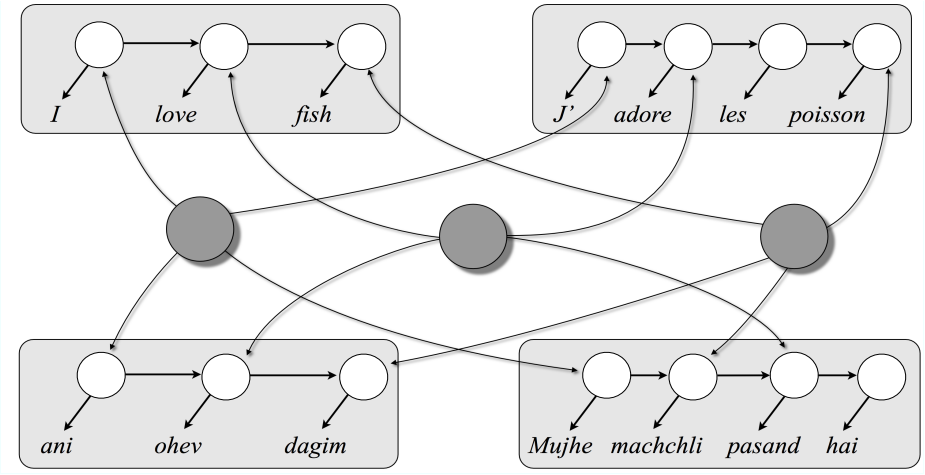

Figure 1: Model structure for parallel sentences in English, French, Hebrew, and Urdu. In this example, there are three superlingual tags, each connected to the part-of-speech tag of a word in each of the four languages.

thus reduce the overall uncertainty in tagging decisions. To encourage the discovery of a compact set of such cross-lingual patterns, we place a Dirichlet process prior on the superlingual tag values.

3.1 Model Structure

For each language, our model includes an HMM-like substructure with observed word nodes, hidden part-of-speech nodes, and directed transition and emission edges. For each set of aligned words in parallel sentences, we add a latent superlingual variable to capture the cross-lingual context. A set of directed edges connect this variable to the part-of-speech nodes of the aligned words. Our model assumes that the superlingual tags for parallel sentences are unordered and are drawn independently of one another.

Edges radiate outward from superlingual tags to language-specific part-of-speech nodes. Thus, our model implicitly assumes that superlingual tags are drawn prior to the part-of-speech tags of all languages and probabilistically influence their selection. See Figure 1 for an example structure.

The particular model structure for each set of parallel sentences (i.e. the configuration of superlingual tags and their edges) is determined by bilingual lexical alignments and — like the text itself — is considered an observed variable. In practice, these lexical alignments are obtained using standard techniques from machine translation.

Our model design has several benefits. Crossing and many-to-many alignments may be used without creating cycles in the graph, as all cross-lingual information emanates from the hidden superlingual tags. Furthermore, the model scales gracefully with the number of languages, as the number of new edges and nodes will be proportional to the number of words for each additional language.

3.2 Superlingual Tags

Each superlingual tag value specifies a set of distributions — one for each language's part-of-speech tagset. In order to learn repeated cross-lingual patterns, we need to constrain the number of superlingual tag values and thus the number of distributions they provide. For example, we might allow the superlingual tags to take on integer values from 1 to K, with each integer value indexing a separate set of distributions. Each set of distributions should correspond to a discovered cross-lingual pattern in the data. For example, one set of distributions might favor nouns in each language and another might favor verbs.

Rather than fixing the number of superlingual tag values to an arbitrary and predetermined size $1, \ldots, K$, we allow them to range over the entire set of integers. In order to encourage the desired multilingual clustering behavior, we use a Dirichlet process prior for the superlingual tags. This prior allows high posterior probability only when a small number

of values are used repeatedly. The actual number of sampled values will be dictated by the data and the number of languages.

More formally, suppose we have n languages, $\ell_1, \ldots, \ell_n$. According to our generative model, a countably infinite sequence of sets $\langle \omega_1^{\ell_1}, \ldots, \omega_1^{\ell_n} \rangle, \langle \omega_2^{\ell_1}, \ldots, \omega_2^{\ell_n} \rangle, \ldots$ is drawn from some base distribution. Each ω_i^{ℓ} is a distribution over the parts-of-speech in language ℓ.

In parallel, an infinite sequence of mixing components $\pi_1, \pi_2, \ldots$ is drawn from a stick-breaking process (Sethuraman, 1994). These components define a distribution over the integers with most probability mass placed on some initial set of values. The two sequences $\langle \omega_1^{\ell_1}, \ldots, \omega_1^{\ell_n} \rangle, \langle \omega_2^{\ell_1}, \ldots, \omega_2^{\ell_n} \rangle \ldots$ and $\pi_1, \pi_2 \ldots$ now define the distribution over superlingual tags and their associated distributions on parts-of-speech. That is, each superlingual tag $z \in \mathbb{N}$ is drawn with probability π_z, and indexes the set of distributions $\langle \omega_z^{\ell_1}, \ldots, \omega_z^{\ell_n} \rangle$.

3.3 Part-of-Speech Tags

Finally, we need to define the generative probabilities of the part-of-speech nodes. For each such node there may be multiple incoming edges. There will always be an incoming transition edge from the previous tag (in the same language). In addition, there may be incoming edges from zero or more superlingual tags. Each edge carries with it a distribution over parts-of-speech and these distributions must be combined into the single distribution from which the tag is ultimately drawn.

We choose to combine these distributions as a product of experts. More formally: for language ℓ and tag position i, the part-of-speech tag y_i is drawn according to

$$y_i \sim \frac{\phi_{y_{i-1}}(y_i) \prod_z \omega_z^{\ell}(y_i)}{Z} \tag{1}$$

Where $\phi_{y_{i-1}}$ indicates the transition distribution, and the z's range over the values of the incoming superlingual tags. The normalization term Z is obtained by summing the numerator over all part-of-speech tags y_i in the tagset.

This parameterization allows for a relatively simple and small parameter space. It also leads to a desirable property: for a tag to have high probability *each* of the incoming distributions must allow it. That is, any expert can "veto" a potential tag by assigning it low probability, generally leading to consensus decisions.

We now formalize this description by giving the stochastic process by which the observed data (raw parallel text) is generated, according to our model.

3.4 Generative Process

For n languages, we assume the existence of n tagsets $T^1, \ldots, T^n$ and vocabularies, $W^1, \ldots, W^n$, one for each language. For clarity, the generative process is described using only bigram transition dependencies, but our experiments use a trigram model.

1. **Transition and Emission Parameters:** For each language ℓ and for each tag $t \in T^{\ell}$, draw a *transition* distribution ϕ_t^{ℓ} over tags T_{ℓ} and an *emission* distribution θ_t^{ℓ} over words W^{ℓ}, all from symmetric Dirichlet priors of appropriate dimension.

2. **Superlingual Tag Parameters:** Draw an infinite sequence of sets $\langle \omega_1^{\ell_1}, \ldots, \omega_1^{\ell_n} \rangle, \langle \omega_2^{\ell_1}, \ldots, \omega_2^{\ell_n} \rangle, \ldots$ from base distribution G_0. Each ω_i^{ℓ} is a distribution over the tagset T^{ℓ}. The base distribution G_0 is a product of n symmetric Dirichlets, where the dimension of the i^{th} such Dirichlet is the size of the corresponding tagset T^{ℓ_i}.

   At the same time, draw an infinite sequence of mixture weights $\pi \sim GEM(\alpha)$, where $GEM(\alpha)$ indicates the stick-breaking distribution (Sethuraman, 1994), and $\alpha = 1$. These parameters together define a prior distribution over superlingual tags,

$$p(z) = \sum_k^{\infty} \pi_k \delta_{k=z}, \tag{2}$$

   or equivalently over the part-of-speech distributions $\langle \omega^{\ell_1}, \ldots, \omega^{\ell_n} \rangle$ that they index:

$$\sum_k^{\infty} \pi_k \delta_{\langle \omega_k^{\ell_1}, \ldots, \omega_k^{\ell_n} \rangle = \langle \omega^{\ell_1}, \ldots, \omega^{\ell_n} \rangle}. \tag{3}$$

   In both cases, $\delta_{v=v'}$ is defined as one when $v = v'$ and zero otherwise. Distribution 3 is said to be drawn from a Dirichlet process, conventionally written as $DP(\alpha, G_0)$.

3. **Data:** For each multilingual parallel sentence,
   (a) Draw an alignment a specifying sets of aligned indices across languages. Each such set may consist of indices in any subset of the languages. We leave the distribution over alignments undefined, as we consider alignments observed variables.
   (b) For each set of indices in a, draw a superlingual tag value z according to Distribution 2.
   (c) For each language ℓ, for $i = 1, \ldots$ (until end-tag reached):
      i. Draw a part-of-speech tag $y_i \in T^\ell$ according to Distribution 1
      ii. Draw a word $w_i \in W^\ell$ according to the emission distribution θ_{y_i}.

To perform Bayesian inference under this model we use a combination of sampling techniques, which we describe in detail in the next section.

3.5 Inference

Ideally we would like to predict the part-of-speech tags which have highest *marginal* probability given the observed words $\mathbf{x}$ and alignments $\mathbf{a}$. More specifically, since we are evaluating our accuracy per tag-position, we would like to predict, for language index ℓ and word index i, the single part-of-speech tag:

$$\operatorname*{argmax}_{t \in T^\ell} P\big(y_i^\ell = t \big| \mathbf{x}, \mathbf{a}\big)$$

which we can rewrite as the $\operatorname{argmax}_{t \in T^\ell}$ of the integral,

$$\int \Bigg[P\left(y_i^\ell = t \Big| \mathbf{y}_{-(\ell,i)}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\omega}, \mathbf{x}, \mathbf{a}\right) \cdot P\left(\mathbf{y}_{-(\ell,i)}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\omega}, \Big| \mathbf{x}, \mathbf{a}\right) \Bigg] d\mathbf{y}_{-(\ell,i)}\, d\boldsymbol{\phi}\, d\boldsymbol{\theta}\, d\mathbf{z}\, d\boldsymbol{\pi}\, d\boldsymbol{\omega},$$

in which we marginalize over the settings of all tags other than y_i^ℓ (written as $\mathbf{y}_{-(\ell,i)}$), the transition distributions $\boldsymbol{\phi} = \left\{\phi_{t'}^\ell\right\}$, emission distributions $\boldsymbol{\theta} = \left\{\theta_{t'}^\ell\right\}$, superlingual tags $\mathbf{z}$, and superlingual tag parameters $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$ and $\boldsymbol{\omega} = \left\{\langle \omega_1^{\ell_1}, \ldots, \omega_1^{\ell_n} \rangle, \langle \omega_2^{\ell_1}, \ldots, \omega_2^{\ell_n} \rangle \ldots\right\}$ (where t' ranges over all part-of-speech tags).

As these integrals are intractable to compute exactly, we resort to the standard Monte Carlo approximation. We collect N samples of the variables over which we wish to marginalize but for which we cannot compute closed-form integrals, where each sample $sample_k$ is drawn from $P(sample_k|\mathbf{x}, \mathbf{a})$. We then approximate the tag marginals as:

$$P\big(y_i^\ell = t \big| \mathbf{x}, \mathbf{a}\big) \approx \frac{\sum_k P\big(y_i^\ell = t \big| sample_k, \mathbf{x}, \mathbf{a}\big)}{N} \quad (4)$$

We employ closed forms for integrating out the emission parameters $\boldsymbol{\theta}$, transition parameters $\boldsymbol{\phi}$, and superlingual tag parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$. We explicitly sample only part-of-speech tags $\mathbf{y}$, superlingual tags $\mathbf{z}$, and the hyperparameters of the transition and emission Dirichlet priors. To do so, we apply standard Markov chain sampling techniques: a Gibbs sampler for the tags and a within-Gibbs Metropolis-Hastings subroutine for the hyperparameters (Hastings, 1970).

Our Gibbs sampler samples each part-of-speech and superlingual tag separately, conditioned on the current value of all other tags. In each case, we use standard closed forms to integrate over all parameter values, using currently sampled counts and hyperparameter pseudo-counts. We note that conjugacy is technically broken by our use of a product form in Distribution 1. Nevertheless, we consider the sampled tags to have been generated separately by each of the factors involved in the numerator. Thus our method of using count-based closed forms should be viewed as an approximation.

3.6 Sampling Part-of-Speech Tags

To sample the part-of-speech tag for language ℓ at position i we draw from

$$P(y_i^\ell | \mathbf{y}_{-(\ell,i)}, \mathbf{x}, \mathbf{a}, \mathbf{z}) \propto P(y_{i+1}^\ell | y_i^\ell, \mathbf{y}_{-(\ell,i)}, \mathbf{a}, \mathbf{z})\; P(y_i^\ell | \mathbf{y}_{-(\ell,i)}, \mathbf{a}, \mathbf{z}) \cdot P(x_i^\ell | \mathbf{x}_{-i}^\ell, \mathbf{y}^\ell)\,,$$

where the first two terms are the generative probabilities of *(i)* the current tag given the previous tag and superlingual tags, and *(ii)* the next tag given the current tag and superlingual tags. These two quantities are similar to Distribution 1, except here we integrate over the transition parameter $\phi_{y_{i-1}}$ and the superlingual tag parameters ω_z^ℓ. We end up with a product of integrals. Each integral can be computed in closed form using multinomial-Dirichlet conjugacy (and by making the above-mentioned simplifying assumption that all other tags were generated separately by their transition and superlingual

parameters), just as in the monolingual Bayesian HMM of (Goldwater and Griffiths, 2007).

For example, the closed form for integrating over the parameter of a superlingual tag with value z is given by:

$$\int \omega_z^\ell(y_i) P(\omega_z^\ell|\omega_0) d\omega_z^\ell = \frac{count(z, y_i, \ell) + \omega_0}{count(z, \ell) + T^\ell \omega_0}$$

where $count(z, y_i, \ell)$ is the number of times that tag y_i is observed together with superlingual tag z in language ℓ, $count(z, \ell)$ is the total number of times that superlingual tag z appears with an edge into language ℓ, and ω_0 is a hyperparameter.

The third term in the sampling formula is the emission probability of the current word x_i^ℓ given the current tag and all other words and sampled tags, as well as a hyperparameter which is suppressed for the sake of clarity. This quantity can be computed exactly in closed form in a similar way.

3.7 Sampling Superlingual Tags

For each set of aligned words in the observed alignment $\mathbf{a}$ we need to sample a superlingual tag z. Recall that z is an index into an infinite sequence $\langle \omega_1^{\ell_1}, \ldots, \omega_1^{\ell_n} \rangle, \langle \omega_2^{\ell_1}, \ldots, \omega_2^{\ell_n} \rangle \ldots$, where each ω_z^ℓ is a distribution over the tagset T^ℓ. The generative distribution over z is given by equation 2. In our sampling scheme, however, we integrate over all possible settings of the mixing components $\boldsymbol{\pi}$ using the standard Chinese Restaurant Process (CRP) closed form (Antoniak, 1974):

$$P(z_i|\mathbf{z}_{-i}, \mathbf{y}) \propto$$
$$\prod_\ell P(y_i^\ell|\mathbf{z}, \mathbf{y}_{-(\ell,i)}) \cdot \begin{cases} \frac{1}{k+\alpha} count(z_i) & \text{if } z_i \in \mathbf{z}_{-i} \\ \frac{\alpha}{k+\alpha} & \text{otherwise} \end{cases}$$

The first term is the product of closed form tag probabilities of the aligned words, given z. The final term is the standard CRP closed form for posterior sampling from a Dirichlet process prior. In this term, k is the total number of sampled superlingual tags, $count(z_i)$ is the total number of times the value z_i occurs in the sampled tags, and α is the Dirichlet process concentration parameter (see Step 2 in Section 3.4).

Finally, we perform standard hyperparameter reestimation for the parameters of the Dirichlet distribution priors on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (the transition and emission distributions) using Metropolis-Hastings. We assume an improper uniform prior and use a Gaussian proposal distribution with mean set to the previous value, and variance to one-tenth of the mean.

4 Experimental Setup

We test our model in an unsupervised framework where only raw parallel text is available for each of the languages. In addition, we assume that for each language a tag dictionary is available that covers some subset of words in the text. The task is to learn an independent tagger for each language that can annotate non-parallel raw text using the learned parameters. All reported results are on non-parallel monolingual test data.

Data For our experiments we use the Multext-East parallel corpus (Erjavec, 2004) which has been used before for multilingual learning (Feldman et al., 2006; Snyder et al., 2008). The tagged portion of the corpus includes a 100,000 word English text, Orwell's novel "Nineteen Eighty Four", and its translation into seven languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene and Serbian. The corpus also includes a tag lexicon for each of these languages. We use the first 3/4 of the text for learning and the last 1/4 as held-out non-parallel test data.

The corpus provides sentence level alignments. To obtain word level alignments, we run GIZA++ (Och and Ney, 2003) on all 28 pairings of the 8 languages. Since we want each latent superlingual variable to span as many languages as possible, we aggregate the pairwise lexical alignments into larger sets of aligned words. These sets of aligned words are generated as a preprocessing step. During sampling they remain fixed and are treated as observed data.

We use the set of 14 basic part-of-speech tags provided by the corpus. In our first experiment, we assume that a complete tag lexicon is available, so that for each word, its set of possible parts-of-speech is known ahead of time. In this setting, the average number of possible tags per token is 1.39. We also experimented with incomplete tag dictionaries, where entries are only available for words appearing more than five or ten times in the corpus. For other words, the entire tagset of 14 tags is considered. In these two scenarios, the average per-token tag ambi-

	Lexicon: Full				Lexicon: Frequency > 5				Lexicon: Frequency > 10			
	MONO	BI		MULTI	MONO	BI		MULTI	MONO	BI		MULTI
		AVG	BEST			AVG	BEST			AVG	BEST	
BG	88.8	91.3	**94.7**	92.6	73.5	80.2	**82.7**	81.3	71.9	77.8	**80.2**	78.8
CS	93.7	97.0	97.7	**98.2**	72.2	79.0	79.7	**83.0**	66.7	75.3	76.7	**79.4**
EN	95.8	95.9	**96.1**	95.0	87.3	90.4	**90.7**	88.1	84.4	88.8	**89.4**	86.1
ET	92.5	93.4	94.3	**94.6**	72.5	76.5	77.5	**80.6**	68.3	72.9	74.9	**77.9**
HU	95.3	96.8	**96.9**	96.7	73.5	77.3	78.0	**80.8**	69.0	73.8	75.2	**76.4**
RO	90.1	91.8	94.0	**95.1**	77.1	82.7	84.4	**86.1**	73.0	80.5	82.1	**83.1**
SL	87.4	89.3	94.8	**95.8**	75.7	78.7	80.9	**83.6**	70.4	76.1	77.6	**80.0**
SR	84.5	90.2	**94.5**	92.3	66.3	75.9	**79.4**	78.8	63.7	72.4	**76.1**	75.9
Avg.	91.0	93.2	**95.4**	95.0	74.7	80.1	81.7	**82.8**	70.9	77.2	79.0	**79.7**

Table 1: Tagging accuracy for Bulgarian, Czech, English, Estonian, Hungarian, Romanian, Slovene, and Serbian. In the first scenario, a complete tag lexicon is available for all the words. In the other two scenarios the tag lexicon only includes words that appear more than five or ten times. Results are given for a monolingual Bayesian HMM (Goldwater and Griffiths, 2007), a bilingual model (Snyder et al., 2008), and the multilingual model presented here. In the case of the bilingual model, we present both the average accuracy over all pairings as well as the result from the best performing pairing for each language. The best results for each language in each scenario are given in boldface.

guity is 4.65 and 5.58, respectively.

Training and testing In the full lexicon experiment, each word is initialized with a random part-of-speech tag from its dictionary entry. In the two reduced lexicon experiments, we initialize the tags with the result of our monolingual baseline (see below) to reduce sampling time. In both cases, we begin with 14 superlingual tag values — corresponding to the parts-of-speech — and initially assign them based on the most common initial part-of-speech of words in each alignment.

We run our Gibbs sampler for 1,000 iterations, and store the conditional tag probabilities for the last 100 iterations. We then approximate marginal tag probabilities on the training data using Equation 4 and predict the highest probability tags. Finally, we compute maximum likelihood transition and emission probabilities using these tag counts, and then apply smoothed viterbi decoding to each held-out monolingual test set. All reported results are averaged over five runs of the sampler.

Monolingual and bilingual baselines We reimplemented the Bayesian HMM model of Goldwater and Griffiths (2007) (BHMM1) as our monolingual baseline. It has a standard HMM structure with conjugate Bayesian priors over transitions and emissions. We note that our model, in the absence of any superlingual tags, reduces to this Bayesian HMM. As an additional baseline we use a bilingual model (Snyder et al., 2008). It is a directed graphical model that jointly tags two parallel streams of text aligned at the word level. The structure of the model consists of two parallel HMMs, one for each language. The aligned words form joint nodes that are shared by both HMMs. These joint nodes are sampled from a probability distribution that is a product of the transition and emission distributions in the two languages and a coupling distribution.

We note that the numbers reported here for the bilingual model differ slightly from those reported by Snyder et al. (2008) for two reasons: we use a slightly larger set of sentences, and an improved sampling scheme. The new sampling scheme marginalizes over the transition and coupling parameters by using the same count-based approximation that we utilize for our multilingual model. This leads to higher performance, and thus a stronger baseline.[1]

5 Results

Table 1 shows the tagging accuracy of our multilingual model on the test data, when training is performed on all eight languages together. Results from both baselines are also reported. In the case of the bilingual baseline, seven pairings are possible for each language, and the results vary by pair. There-

[1] Another difference is that we use the English lexicon provided with the Multext-East corpus, whereas (Snyder et al., 2008) augment this lexicon with tags found in WSJ.

fore, for each language, we present the average accuracy over all seven pairings, as well as the accuracy of its highest performing pairing.

We provide results for three scenarios. In the first case, a tag dictionary is provided for all words, limiting them to a restricted set of possible tags. In the other two scenarios, dictionary entries are limited to words that appear more than five or ten times in the corpus. All other words can be assigned any tag, increasing the overall difficulty of the task. In the full lexicon scenario, our model achieves an average tagging accuracy of 95%, compared to 91% for the monolingual baseline and 93.2% for the bilingual baseline when averaged over all pairings. This accuracy (95%) comes close to the performance of the bilingual model when the best pairing for each language is chosen by an oracle (95.4%). This demonstrates that our multilingual model is able to effectively learn from all languages. In the two reduced lexicon scenarios, the gains are even more striking. In both cases the average multilingual performance outpaces even the *best* performing pairs.

Looking at individual languages, we see that in all three scenarios, Czech, Estonian, Romanian, and Slovene show their best performance with the multilingual model. Bulgarian and Serbian, on the other hand, give somewhat better performance with their optimal pairings under the bilingual model, but their multilingual performance remains higher than their average bilingual results. The performance of English under the multilingual model is somewhat lower, especially in the full lexicon scenario, where it drops below monolingual performance. One possible explanation for this decrease lies in the fact that English, by far, has the lowest trigram tag entropy of all eight languages (Snyder et al., 2008). It is possible, therefore, that the signal it should be getting from its own transitions is being drowned out by less reliable information from other languages.

In order to test the performance of our model as the number of languages increases, we ran the full lexicon experiment with all possible subsets of the eight languages. Figure 2 plots the average accuracy as the number of languages varies. For comparison, the monolingual and average bilingual baseline results are given, along with supervised monolingual performance. Our multilingual model steadily gains in accuracy as the number of available languages increases. Interestingly, it even outperforms the bilingual baseline (by a small margin) when only two languages are available, which may be attributable to the more flexible non-parametric dependencies employed here. Finally, notice that the gap between monolingual supervised and unsupervised performance is cut by nearly two thirds under the unsupervised multilingual model.

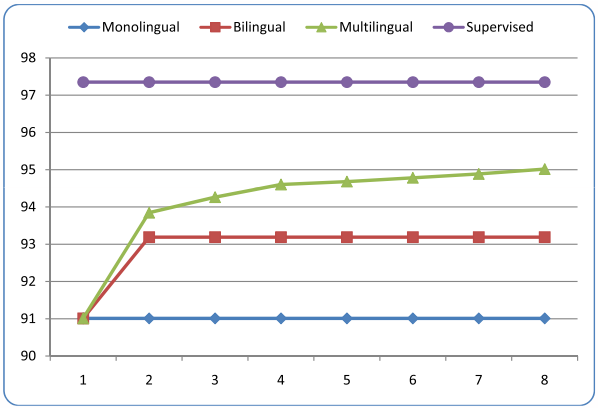

Figure 2: Performance of the multilingual model as the number of languages varies. Performance of the monolingual and average bilingual baselines as well as a supervised monolingual performance are given for comparison.

6 Conclusion

In this paper we've demonstrated that the benefits of unsupervised multilingual learning increase steadily with the number of available languages. Our model scales gracefully as languages are added and effectively incorporates information from them all, leading to substantial performance gains. In one experiment, we cut the gap between unsupervised and supervised performance by nearly two thirds. A future challenge lies in incorporating constraints from additional languages even when parallel text is unavailable.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (CAREER grant IIS-0448168 and grant IIS-0835445). Thanks to Tommi Jaakkola and members of the MIT NLP group for helpful discussions. Any opinions, findings, or recommendations expressed above are those of the authors and do not necessarily reflect the views of the NSF.

References

C. E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, November.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*.

T. Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC*, volume 4, pages 1535–1538.

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.

Dmitriy Genzel. 2005. Inducing a multilingual dictionary from a parallel multitext in related languages. In *Proceedings of the 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.

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

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *MT Summit 2001*, pages 253–258.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

J. Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the EMNLP*, pages 1041–1050.

Masao Utiyama and Hitoshi Isahara. 2006. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL/HLT*, pages 484–491.

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.