Unsupervised Consonant-Vowel Prediction over Hundreds of Languages

Young-Bum Kim and Benjamin Snyder
University of Wisconsin-Madison
{ybkim,bsnyder}@cs.wisc.edu

Abstract

In this paper, we present a solution to one aspect of the decipherment task: the prediction of consonants and vowels for an unknown language and alphabet. Adopting a classical Bayesian perspective, we performs posterior inference over hundreds of languages, leveraging knowledge of known languages and alphabets to uncover general linguistic patterns of typologically coherent language clusters. We achieve average accuracy in the unsupervised consonant/vowel prediction task of 99% across 503 languages. We further show that our methodology can be used to predict more fine-grained phonetic distinctions. On a three-way classification task between vowels, nasals, and non-nasal consonants, our model yields unsupervised accuracy of 89% across the same set of languages.

1 Introduction

Over the past centuries, dozens of lost languages have been deciphered through the painstaking work of scholars, often after decades of slow progress and dead ends. However, several important writing systems and languages remain undeciphered to this day.

In this paper, we present a successful solution to one aspect of the decipherment puzzle: automatically identifying basic phonetic properties of letters in an unknown alphabetic writing system. Our key idea is to use knowledge of the phonetic regularities encoded in known language vocabularies to automatically build a universal probabilistic model to successfully decode new languages.

Our approach adopts a classical Bayesian perspective. We assume that each language has an unobserved set of parameters explaining its observed vocabulary. We further assume that each language-specific set of parameters was itself drawn from an unobserved common prior, shared across a cluster of typologically related languages. In turn, each cluster derives its parameters from a universal prior common to all language groups. This approach allows us to mix together data from languages with various levels of observations and perform joint posterior inference over unobserved variables of interest.

At the bottom layer (see Figure 1), our model assumes a language-specific data generating HMM over words in the language vocabulary. Each word is modeled as an emitted sequence of characters, depending on a corresponding Markov sequence of phonetic tags. Since individual letters are highly constrained in their range of phonetic values, we make the assumption of one-tag-per-observation-type (e.g. a single letter is constrained to be always a consonant or always a vowel across all words in a language).

Going one layer up, we posit that the language-specific HMM parameters are themselves drawn from informative, non-symmetric distributions representing a typologically coherent language grouping. By applying the model to a mix of languages with observed and unobserved phonetic sequences, the cluster-level distributions can be inferred and help guide prediction for unknown languages and alphabets.

We apply this approach to two small decipherment tasks:

1. predicting whether individual characters in an unknown alphabet and language represent vowels or consonants, and

2. predicting whether individual characters in an unknown alphabet and language represent vowels, nasals, or non-nasal consonants.

For both tasks, our approach yields considerable

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1527–1536,
Sofia, Bulgaria, August 4-9 2013.

success. We experiment with a data set consisting of vocabularies of 503 languages from around the world, written in a mix of Latin, Cyrillic, and Greek alphabets. In turn for each language, we consider it and its alphabet "unobserved" — we hide the graphic and phonetic properties of the symbols — while treating the vocabularies of the remaining languages as fully observed with phonetic tags on each of the letters.

On average, over these 503 leave-one-language-out scenarios, our model predicts consonant/vowel distinctions with 99% accuracy. In the more challenging task of vowel/nasal/non-nasal prediction, our model achieves average accuracy over 89%.

2 Related Work

The most direct precedent to the present work is a section in Knight et al. (2006) on universal phonetic decipherment. They build a trigram HMM with three hidden states, corresponding to consonants, vowels, and spaces. As in our model, individual characters are treated as the observed emissions of the hidden states. In contrast to the present work, they allow letters to be emitted by multiple states.

Their experiments show that the HMM trained with EM successfully clusters Spanish letters into consonants and vowels. They further design a more sophisticated finite-state model, based on linguistic universals regarding syllable structure and sonority. Experiments with the second model indicate that it can distinguish sonorous consonants (such as n, m, l, r) from non-sonorous consonants in Spanish. An advantage of the linguistically structured model is that its predictions do not require an additional mapping step from uninterpreted hidden states to linguistic categories, as they do with the HMM.

Our model and experiments can be viewed as complementary to the work of Knight et al., while also extending it to hundreds of languages. We use the simple HMM with EM as our baseline. In lieu of a linguistically designed model structure, we choose an empirical approach, allowing posterior inference over hundreds of known languages to guide the model's decisions for the unknown script and language.

In this sense, our model bears some similarity to the decipherment model of Snyder et al. (2010), which used knowledge of a related language (Hebrew) in an elaborate Bayesian framework to decipher the ancient language of Ugaritic. While the aim of the present work is more modest (discovering very basic phonetic properties of letters) it is also more widely applicable, as we don't required detailed analysis of a known related language.

Other recent work has employed a similar perspective for tying learning across languages. Naseem et al. (2009) use a non-parametric Bayesian model over parallel text to jointly learn part-of-speech taggers across 8 languages, while Cohen and Smith (2009) develop a shared logistic normal prior to couple multilingual learning even in the absence of parallel text. In similar veins, Berg-Kirkpatrick and Klein (2010) develop hierarchically tied grammar priors over languages within the same family, and Bouchard-Côté et al. (2013) develop a probabilistic model of sound change using data from 637 Austronesian languages.

In our own previous work, we have developed the idea that *supervised* knowledge of some number of languages can help guide the unsupervised induction of linguistic structure, even in the absence of parallel text (Kim et al., 2011; Kim and Snyder, 2012)[1]. In the latter work we also tackled the problem of unsupervised phonemic prediction for unknown languages by using textual regularities of known languages. However, we assumed that the target language was written in a known (Latin) alphabet, greatly reducing the difficulty of the prediction task. In our present case, we assume no knowledge of any relationship between the writing system of the target language and known languages, other than that they are all alphabetic in nature.

Finally, we note some similarities of our model to some ideas proposed in other contexts. We make the assumption that each observation type (letter) occurs with only one hidden state (consonant or vowel). Similar constraints have been developed for part-of-speech tagging (Lee et al., 2010; Christodoulopoulos et al., 2011), and the power of type-based sampling has been demonstrated, even in the absence of explicit model constraints (Liang et al., 2010).

3 Model

Our generative Bayesian model over the observed vocabularies of hundreds of languages is

[1] We note that similar ideas were simultaneously proposed by other researchers (Cohen et al., 2011).

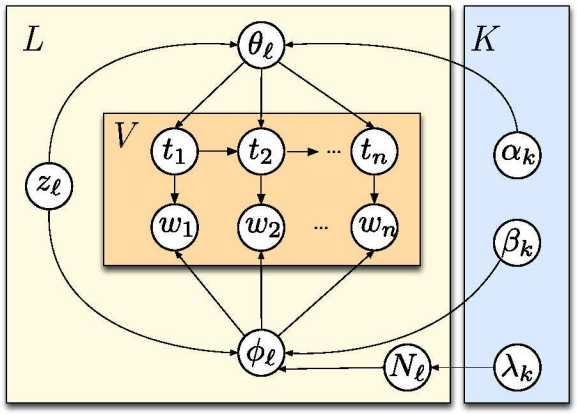

Figure 1: Graphical representation of our model. We have K language clusters, L languages, and V words in each language.

presented in Figure 1 and Algorithms 1, 2, and 3. We present a running commentary on the generative process from the bottom up, starting with Algorithm 3.

3.1 Data Generation

At the data generation stage (Algorithm 3), our model resembles an HMM. At each time step i, a tag t_i is selected according to a language-specific transition distribution ϕ, indexed by the previous tag t_{i-1}. We note that in practice, we implemented a trigram version of the model,[2] but we present the bigram version here for notational clarity. We assume that our tagset includes phonetic categories of interest (such as consonant, vowel, nasal, etc) as well as a special tag to denote the boundaries between words.

An observation index $j \in 1 \ldots N_{\ell,t_i}$ is then drawn from the language-specific emission distribution ϕ, indexed by the current tag t_i. N_{ℓ,t_i} denotes the number of observation types associated with tag t_i in language ℓ. Finally, we assume the existence of a deterministic function `orth` which maps each tag's observation indices to unique orthographic character symbols. This ensures that each observed character type corresponds to an observation index in exactly one tag category.

3.2 Language Generation

At the next stage up (Algorithm 2), we consider the generation of all language-specific parameters. This process begins by selecting a language cluster assignment z_ℓ uniformly. The language cluster

[2] where the transition distribution is indexed by the previous *two* tags

Algorithm 1 Cluster Generation

for each cluster $k \in 1 \ldots K$ **do**
  for each tag $t \in 1 \ldots T$ **do**
    // emission Dirichlet parameter
    $\beta_{k,t} \sim \text{Unif}[0, 500]$
    // type-count Poisson parameter
    $\lambda_{k,t} \sim \text{Gamma}(g_1, g_2)$
    // transition Dirichlet parameters
    for each tag t' **do**
      $\alpha_{k,t,t'} \sim \text{Unif}[0, 500]$

Algorithm 2 Language Generation

for each language ℓ **do**
  // draw cluster assignment
  cluster $z_\ell \sim \text{Unif}[1 \ldots K]$
  for each tag t **do**
    // generate tag type-count
    $N_{\ell,t} \sim \text{Poisson}(\lambda_{z_\ell,t})$
    // generate emission multinomial
    $\phi_{\ell,t,1} \ldots \phi_{\ell,t,N_{\ell,t}} \sim \text{SymmDir}(\beta_{z_\ell,t})$
    // generate transition multinomial
    $\theta_{\ell,t,1} \ldots \theta_{\ell,t,T} \sim \text{Dir}(\alpha_{z_\ell,t,1} \ldots \alpha_{z_\ell,t,T})$

Algorithm 3 Data Generation

for each language ℓ **do**
  for each position i **do**
    // transition to new tag token
    $t_i | t_{i-1} \sim \text{Cat}(\theta_{\ell,t_{i-1},1} \ldots \theta_{\ell,t_{i-1},T})$
    // emit observation index token
    $j | t_i \sim \text{Cat}(\phi_{\ell,t_i,1} \ldots \phi_{\ell,t_i,N_{\ell,t_i}})$
    // transcribe index token as character
    $w_i \leftarrow \texttt{orth}(\ell, j, t_i)$

provides priors over the HMM parameters. These priors include:

1. Poisson distributions over the number of observation types $N_{\ell,t}$ associated with tag t,
2. Dirichlet priors over transition distributions θ, and
3. Dirichlet priors over emission distributions ϕ.

For example, the cluster Poisson parameter over vowel observation types might be $\lambda = 9$ (indicating 9 vowel letters on average for the cluster), while the parameter over consonant observation types might be $\lambda = 20$ (indicating 20 consonant letters on average). These priors will be distinct for each language cluster and serve to characterize its general linguistic and typological properties.

We pause at this point to review the Dirichlet distribution in more detail. A $k-$dimensional Dirichlet with parameters $\alpha_1 ... \alpha_k$ defines a distribution over the $k-1$ simplex with the following density:

$$f(\theta_1 ... \theta_k | \alpha_1 ... \alpha_k) \propto \prod_i \theta_i^{\alpha_i - 1}$$

where $\alpha_i > 0, \theta_i > 0$, and $\sum_i \theta_i = 1$. The Dirichlet serves as the *conjugate prior* for the Multinomial, meaning that the posterior $\theta_1 ... \theta_k | X_1 ... X_n$ is again distributed as a Dirichlet (with updated parameters). It is instructive to reparameterize the Dirichlet with $k+1$ parameters:

$$f(\theta_1 ... \theta_k | \alpha_0, \alpha'_1 ... \alpha'_k) \propto \prod_i \theta_i^{\alpha_0 \alpha'_i - 1}$$

where $\alpha_0 = \sum_i \alpha_i$, and $\alpha'_i = \alpha_i / \alpha_0$. In this parameterization, we have $\mathbb{E}[\theta_i] = \alpha'_i$. In other words, the parameters α'_i give the mean of the distribution, and α_0 gives the *precision* of the distribution. For large $\alpha_0 \gg k$, the distribution is highly peaked around the mean (conversely, when $\alpha_0 \ll k$, the mean lies in a valley).

Thus, the Dirichlet parameters of a language cluster characterize both the average HMMs of individual languages within the cluster, as well as how much we expect the HMMs to vary from the mean. In the case of emission distributions, we assume symmetric Dirichlet priors — i.e. one-parameter Dirichlets with densities given by $f(\theta_1 ... \theta_k | \beta) \propto \prod \theta_i^{(\beta - 1)}$. This assumption is necessary, as we have no way to identify characters across languages in the decipherment scenario, and even the number of consonants and vowels (and thus multinomial/Dirichlet dimensions) can vary across the languages of a cluster. Thus, the mean of these Dirichlets will always be a uniform emission distribution. The single Dirichlet emission parameter per cluster will specify whether this mean is on a peak (large β) or in a valley (small β). In other words, it will control the expected *sparsity* of the resulting per-language emission multinomials.

In contrast, the transition Dirichlet parameters may be asymmetric, and thus very specific and informative. For example, one cluster may have the property that CCC consonant clusters are exceedingly rare across all its languages. This property would be expressed by a very small mean $\alpha'_{\text{CCC}} \ll 1$ but large precision α_0. Later we shall see examples of learned transition Dirichlet parameters.

3.3 Cluster Generation

The generation of the cluster parameters (Algorithm 1) defines the highest layer of priors for our model. As Dirichlets lack a standard conjugate prior, we simply use uniform priors over the interval $[0, 500]$. For the cluster Poisson parameters, we use conjugate Gamma distributions with vague priors.[3]

4 Inference

In this section we detail the inference procedure we followed to make predictions under our model. We run the procedure over data from 503 languages, assuming that all languages but one have observed character and tag sequences: $w_1, w_2, \ldots, t_1, t_2, \ldots$ Since each character type w is assumed to have a single tag category, this is equivalent to observing the character token sequence along with a character-type-to-tag mapping t_w. For the target language, we observe only character token sequence $w_1, w_2, \ldots$

We assume fixed and known parameter values only at the cluster generation level. Unobserved variables include (i) the cluster parameters α, β, λ, (ii) the cluster assignments $\mathbf{z}$, (iii) the per-language HMM parameters θ, ϕ for all languages, and (iv) for the target language, the tag tokens $t_1, t_2, \ldots$ — or equivalently the character-type-to-tag mappings t_w — along with the observation type-counts N_t.

4.1 Monte Carlo Approximation

Our goal in inference is to predict the most likely tag $t_{w,\ell}$ for each character type w in our target language ℓ according to the posterior:

$$f\left(t_{w,\ell} \mid \mathbf{w}, \mathbf{t}_{-\ell}\right) = \int f\left(\mathbf{t}_\ell, \mathbf{z}, \alpha, \beta \mid \mathbf{w}, \mathbf{t}_{-\ell}\right) d\,\Theta \quad (1)$$

[3] (1,19) for consonants, (1,10) for vowels, (0.2, 15) for nasals, and (1,16) for non-nasal consonants.

where $\Theta = (\mathbf{t}_{-w,\ell}, \mathbf{z}, \alpha, \beta)$, $\mathbf{w}$ are the observed character sequences for all languages, $\mathbf{t}_{-\ell}$ are the character-to-tag mappings for the observed languages, $\mathbf{z}$ are the language-to-cluster assignments, and α and β are all the cluster-level transition and emission Dirichlet parameters.

Sampling values $(\mathbf{t}_\ell, \mathbf{z}, \alpha, \beta)_{n=1}^{N}$ from the integrand in Equation 1 allows us to perform the standard Monte Carlo approximation:

$$f\left(t_{w,\ell} = t \mid \mathbf{w}, \mathbf{t}_{-\ell}\right) \approx N^{-1} \sum_{n=1}^{N} \mathbb{I}\left(t_{w,\ell} = t \text{ in sample } n\right) \quad (2)$$

To maximize the Monte Carlo posterior, we simply take the most commonly sampled tag value for character type w in language ℓ. Note that we leave out the language-level HMM parameters (θ, ϕ) as well as the cluster-level Poisson parameters λ from Equation 1 (and thus our sample space), as we can analytically integrate them out in our sampling equations.

4.2 Gibbs Sampling

To sample values $(\mathbf{t}_\ell, \mathbf{z}, \alpha, \beta)$ from their posterior (the integrand of Equation 1), we use Gibbs sampling, a Monte Carlo technique that constructs a Markov chain over a high-dimensional sample space by iteratively sampling each variable conditioned on the currently drawn sample values for the others, starting from a random initialization. The Markov chain converges to an equilibrium distribution which is in fact the desired joint density (Geman and Geman, 1984). We now sketch the sampling equations for each of our sampled variables.

Sampling $t_{w,\ell}$

To sample the tag assignment to character w in language ℓ, we need to compute:

$$f\left(t_{w,\ell} \mid \mathbf{w}, \mathbf{t}_{-w,\ell}, \mathbf{t}_{-\ell}, \mathbf{z}, \alpha, \beta\right) \quad (3)$$
$$\propto f\left(\mathbf{w}_\ell, \mathbf{t}_\ell, N_\ell \mid \alpha_k, \beta_k, \mathbf{N}_{k-\ell}\right) \quad (4)$$

where N_ℓ are the types-per-tag counts implied by the mapping $\mathbf{t}_\ell$, k is the current cluster assignment for the target language ($z_\ell = k$), α_k and β_k are the cluster parameters, and $\mathbf{N}_{k-\ell}$ are the types-per-tag counts for all languages currently assigned to the cluster, *other* than language ℓ.

Applying the chain rule along with our model's conditional independence structure, we can further re-write Equation 4 as a product of three terms:

$$f(N_\ell | \mathbf{N}_{k-\ell}) \quad (5)$$
$$f(t_1, t_2, \ldots | \alpha_k) \quad (6)$$
$$f(w_1, w_2, \ldots | N_\ell, t_1, t_2, \ldots, \beta_k) \quad (7)$$

The first term is the posterior predictive distribution for the Poisson-Gamma compound distribution and is easy to derive. The second term is the tag transition predictive distribution given Dirichlet hyperparameters, yielding a familiar Polya urn scheme form. Removing terms that don't depend on the tag assignment $t_{\ell,w}$ gives us:

$$\frac{\prod_{t,t'} \left(\alpha_{k,t,t'} + n(t,t')\right)^{[n'(t,t')]}}{\prod_{t} \left(\sum_{t'} \alpha_{k,t,t'} + n(t)\right)^{[n'(t)]}}$$

where $n(t)$ and $n(t,t')$ are, respectively, unigram and bigram tag counts *excluding* those containing character w. Conversely, $n'(t)$ and $n'(t,t')$ are, respectively, unigram and bigram tag counts *only including* those containing character w. The notation $a^{[n]}$ denotes the ascending factorial: $a(a+1)\cdots(a+n-1)$. Finally, we tackle the third term, Equation 7, corresponding to the predictive distribution of emission observations given Dirichlet hyperparameters. Again, removing constant terms gives us:

$$\frac{\beta_{k,t}^{[n(w)]}}{\prod_{t'} N_{\ell,t'} \beta_{k,t'}^{[n(t')]}}$$

where $n(w)$ is the unigram count of character w, and $n(t')$ is the unigram count of tag t, over all characters tokens (including w).

Sampling $\alpha_{k,t,t'}$

To sample the Dirichlet hyperparameter for cluster k and transition $t \to t'$, we need to compute:

$$\begin{aligned} &f(\alpha_{k,t,t'} | \mathbf{t}, \mathbf{z}) \\ &\propto f(\mathbf{t}, \mathbf{z} | \alpha_{z,t,t'}) \\ &= f(\mathbf{t}_k | \alpha_{z,t,t'}) \end{aligned}$$

where $\mathbf{t}_k$ are the tag sequences for all languages currently assigned to cluster k. This term is a predictive distribution of the multinomial-Dirichlet compound when the observations are grouped into *multiple* multinomials all with the same prior. Rather than inefficiently computing a product of Polya urn schemes (with many repeated ascending

factorials with the same base), we group common terms together and calculate:

$$\frac{\prod_{j=1}(\alpha_{k,t,t'}+k)^{n(j,k,t,t')}}{\prod_{j=1}(\sum_{t''}\alpha_{k,t,t''}+k)^{n(j,k,t)}}$$

where $n(j,k,t)$ and $n(j,k,t,t')$ are the numbers of languages currently assigned to cluster k which have *more than* j occurrences of unigram (t) and bigram (t,t'), respectively.

This gives us an efficient way to compute unnormalized posterior densities for α. However, we need to sample from these distributions, not just compute them. To do so, we turn to slice sampling (Neal, 2003), a simple yet effective auxiliary variable scheme for sampling values from unnormalized but otherwise computable densities.

The key idea is to supplement the variable x, distributed according to unnormalized density $\tilde{p}(x)$, with a second variable u with joint density defined as $p(x,u) \propto \mathbb{I}(u < \tilde{p}(x))$. It is easy to see that $\tilde{p}(x) \propto \int p(x,u)du$. We then iteratively sample $u|x$ and $x|u$, both of which are distributed uniformly across appropriately bounded intervals. Our implementation follows the pseudocode given in Mackay (2003).

Sampling $\beta_{k,t}$

To sample the Dirichlet hyperparameter for cluster k and tag t we need to compute:

$$\begin{aligned} &f(\beta_{k,t}|\mathbf{t},\mathbf{w},\mathbf{z},\mathbf{N}) \\ &\propto f(\mathbf{w}|\mathbf{t},\mathbf{z},\beta_{k,t},\mathbf{N}) \\ &\propto f(\mathbf{w}_k|\mathbf{t}_k,\beta_{k,t},\mathbf{N}_k) \end{aligned}$$

where, as before, $\mathbf{t}_k$ are the tag sequences for languages assigned to cluster k, $\mathbf{N}_k$ are the tag observation type-counts for languages assigned to the cluster, and likewise $\mathbf{w}_k$ are the character sequences of all languages in the cluster. Again, we have the predictive distribution of the multinomial-Dirichlet compound with multiple grouped observations. We can apply the same trick as above to group terms in the ascending factorials for efficient computation. As before, we use slice sampling for obtaining samples.

Sampling z_ℓ

Finally, we consider sampling the cluster assignment z_ℓ for each language ℓ. We calculate:

$$\begin{aligned} &f(z_\ell = k|\mathbf{w},\mathbf{t},\mathbf{N},\mathbf{z}_{-\ell},\alpha,\beta) \\ &\propto f(\mathbf{w}_\ell,\mathbf{t}_\ell,N_\ell|\alpha_k,\beta_k,\mathbf{N}_{k-\ell}) \\ &= f(N_\ell|\mathbf{N}_{k-\ell})f(\mathbf{t}_\ell|\alpha_k)f(\mathbf{w}_\ell|\mathbf{t}_\ell,N_\ell,\beta_k) \end{aligned}$$

The three terms correspond to (1) a standard predictive distributions for the Poisson-gamma compound and (2) the standard predictive distributions for the transition and emission multinomial-Dirichlet compounds.

5 Experiments

To test our model, we apply it to a corpus of 503 languages for two decipherment tasks. In both cases, we will assume no knowledge of our target language or its writing system, other than that it is alphabetic in nature. At the same time, we will assume basic phonetic knowledge of the writing systems of the other 502 languages. For our first task, we will predict whether each character type is a consonant or a vowel. In the second task, we further subdivide the consonants into two major categories: the nasal consonants, and the non-nasal consonants. Nasal consonants are known to be perceptually very salient and are unique in being high frequency consonants in all known languages.

5.1 Data

Our data is drawn from online electronic translations of the Bible (`http://www.bible.is`, `http://www.crosswire.org/index.jsp`, and `http://www.biblegateway.com`). We have identified translations covering 503 distinct languages employing alphabetic writing systems. Most of these languages (476) use variants of the Latin alphabet, a few (26) use Cyrillic, and one uses the Greek alphabet. As Table 1 indicates, the languages cover a very diverse set of families and geographic regions, with Niger-Congo languages being the largest represented family.[4] Of these languages, 30 are either language isolates, or sole members of their language family in our data set.

For our experiments, we extracted unique word types occurring at least 5 times from the downloaded Bible texts. We manually identified vowel, nasal, and non-nasal character types. Since the letter "y" can frequently represent both a consonant and vowel, we exclude it from our evaluation. On average, the resulting vocabularies contain 2,388 unique words, with 19 consonant characters, two 2 nasal characters, and 9 vowels. We include the data as part of the paper.

[4] In fact, the Niger-Congo grouping is often considered the largest language family in the world in terms of distinct member languages.

Language Family	#lang
Niger-Congo	114
Austronesian	67
Oto-Manguean	41
Indo-European	39
Mayan	34
Quechuan	17
Afro-Asiatic	17
Uto-Aztecan	16
Altaic	16
Trans-New Guinea	15
Nilo-Saharan	14
Sino-Tibetan	13
Tucanoan	9
Creole	8
Chibchan	6
Maipurean	5
Tupian	5
Nakh-Daghestanian	4
Uralic	4
Cariban	4
Totonacan	4
Mixe-Zoque	3
Jivaroan	3
Choco	3
Guajiboan	2
Huavean	2
Austro-Asiatic	2
Witotoan	2
Jean	2
Paezan	2
Other	30

Table 1: Language families in our data set. The *Other* category includes 9 language isolates and 21 language family singletons.

5.2 Baselines and Model Variants

As our baseline, we consider the trigram HMM model of Knight et al. (2006), trained with EM. In all experiments, we run 10 random restarts of EM, and pick the prediction with highest likelihood. We map the induced tags to the gold-standard tag categories (1-1 mapping) in the way that maximizes accuracy.

We then consider three variants of our model. The simplest version, SYMM, disregards all information from other languages, using simple symmetric hyperparameters on the transition and emission Dirichlet priors (all hyperparameters set to 1). This allows us to assess the performance of our Gibbs sampling inference method for the type-based HMM, even in the absence of multilingual priors.

We next consider a variant of our model, MERGE, that assumes that *all* languages reside in a single cluster. This allows knowledge from the other languages to affect our tag posteriors in a generic, language-neutral way.

Finally, we consider the full version of our model, CLUST, with 20 language clusters. By allowing for the division of languages into smaller groupings, we hope to learn more specific parameters tailored for typologically coherent clusters of languages.

	Model	Cons vs Vowel	C vs V vs N
All	EM	93.37	74.59
	SYMM	95.99	80.72
	MERGE	97.14	86.13
	CLUST	**98.85**	**89.37**
Isolates	EM	94.50	74.53
	SYMM	96.18	78.13
	MERGE	97.66	86.47
	CLUST	**98.55**	**89.07**
Non-Latin	EM	92.93	78.26
	SYMM	95.90	79.04
	MERGE	96.06	83.78
	CLUST	**97.03**	**85.79**

Table 2: Average accuracy for EM baseline and model variants across 503 languages. **First panel:** results on all languages. **Second panel:** results for 30 isolate and singleton languages. **Third panel:** results for 27 non-Latin alphabet languages (Cyrillic and Greek). Standard Deviations across languages are about 2%.

6 Results

The results of our experiments are shown in Table 2. In all cases, we report token-level accuracy (i.e. frequent characters count more than infrequent characters), and results are macro-averaged over the 503 languages. Variance across languages is quite low: the standard deviations are about 2 percentage points.

For the consonant vs. vowel prediction task, all tested models perform well. Our baseline, the EM-based HMM, achieves 93.4% accuracy. Simply using our Gibbs sampler with symmetric priors boosts the performance up to 96%. Performance

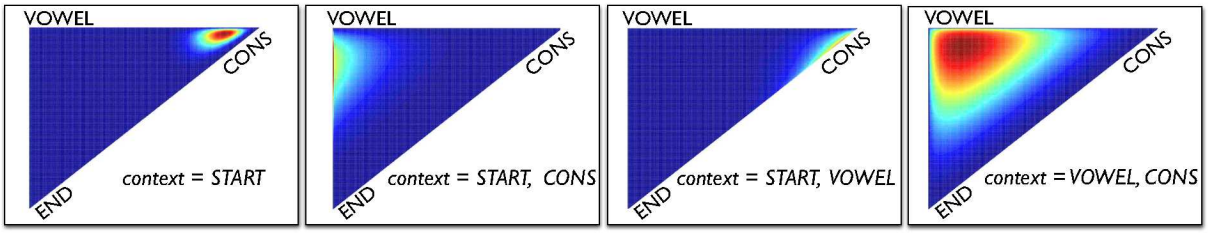

Figure 2: Inferred transition Dirichlet distributions for trigram MERGE model. Heat plots indicate Dirichlet densities over the 2-simplex.

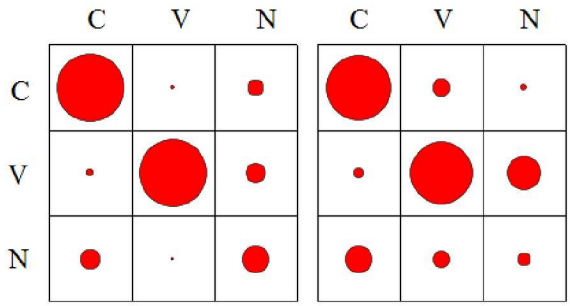

Figure 3: Confusion matrix for CLUST (left) and EM (right). Rows show true values, columns show predicted values. Size of blobs are proportional to counts.

increases again when we condition on other languages (MERGE), and we observe nearly 99% accuracy when allowing languages to cluster.

In the three-way nasal vs. non-nasal consonant vs. vowel prediction task, EM does not fare particularly well, only achieving 75% accuracy. As before, we see increasing performance gains for our model variants, culminating in almost 90% accuracy when the language clustering is used. The relatively weak performance of EM in this case should not be surprising: there is no *a priori* reason to expect any particular three-way classification to be the most salient clustering of letters from the perspective of EM. In contrast, our empirical multilingual approach allows the language-specific tag predictions to be guided by whatever values are set for the other, observed, languages.

We note that although a post-hoc mapping from inferred tags to true tags is necessary for both EM and SYMM, this is not the case for the final two variants of our model. Both MERGE and CLUST break symmetries over tags by way of the asymmetric posterior over transition Dirichlet parameters. Thus the reported accuracies are obtained without the need for any additional tag mappings.

Figure 2 further breaks down results for languages without any other related language in our collection. These include 9 language isolates and 21 singleton languages acting as sole representatives of their families. In addition, we show results for the 27 languages which employ non-Latin alphabets (26 Cyrillic and one Greek). Both of these scenarios are likely to occur in cases of lost language decipherment. We see similar results and trends, with somewhat lower performance in both cases.

7 Analysis

To further compare our model to the EM baseline, we show confusion matrices for the three-way classification task in Figure 3. We can immediately see that EM had considerable difficulty making nasal predictions. Most true nasals (third row) are assigned to the regular consonant category, and apparently EM mostly used the additional tag as a way to further subcategorize vowels. In contrast, our model does fairly well with nasals: most actual nasals are assigned to the nasal category (third row), while the plurality of nasal predictions are indeed true nasals (third column).

Next we examine the transition Dirichlet hyperparameters learned by our model. For the MERGE model, we infer a posterior over parameters shared by all 503 languages in our data set. Figure 2 shows MAP estimates of four of the Dirichlets governing transition probabilities from various contexts. As we can see, the learned hyperparameters yield highly asymmetric priors over transition distributions. Most languages like to start words with consonants, and after an initial consonant or vowel prefer to switch to the opposite category. In contrast, after a vowel-consonant sequence, languages can vary significantly in terms of the category favored next.

Figure 4 shows MAP transition Dirichlet hyperparameters of the CLUST model, when trained

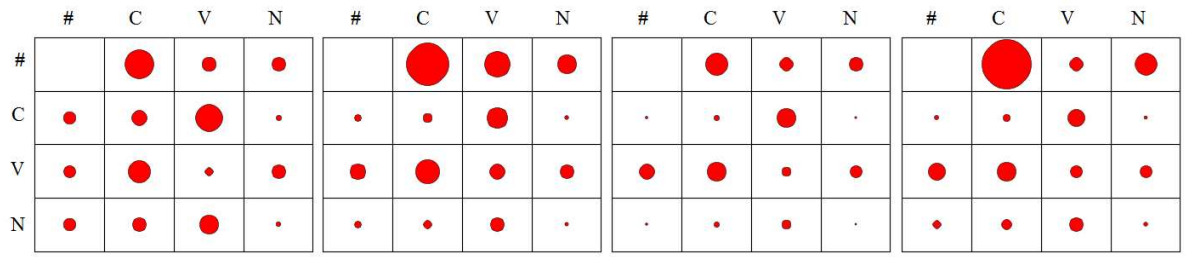

Figure 4: Inferred Dirichlet transition hyperparameters for bigram CLUST on three-way classification task with four latent clusters. Row gives starting state, column gives target state. Size of red blobs are proportional to magnitude of corresponding hyperparameters.

Language Family	Portion	#langs	Ent.
Indo-European	0.38	26	2.26
	0.24	41	3.19
	0.21	38	3.77
Quechuan	0.89	18	0.61
Mayan	0.64	33	1.70
Oto-Manguean	0.55	31	1.99
Maipurean	0.25	8	2.75
Tucanoan	0.2	45	3.98
Uto-Aztecan	0.4	25	2.85
Altaic	0.44	27	2.76
Niger-Congo	1	2	0.00
	0.78	23	1.26
	0.74	27	1.05
	0.68	22	1.22
	0.67	33	1.62
	0.5	18	2.21
	0.24	25	3.27
Austronesian	0.91	22	0.53
	0.71	21	1.51
	0.24	17	3.06

Table 3: Plurality language families across 20 clusters. The columns indicate portion of languages in the plurality family, number of languages, and entropy over families.

with a bigram HMM with four language clusters. Examining just the first row, we see that the languages are partially grouped by their preference for the initial tag of words. All clusters favor languages which prefer initial consonants, though this preference is most weakly expressed in cluster 3. In contrast, both clusters 2 and 4 have very dominant tendencies towards consonant-initial languages, but differ in the relative weight given to languages preferring either vowels or nasals initially.

Finally, we examine the relationship between the induced clusters and language families in Table 3, for the trigram consonant vs. vowel CLUST model with 20 clusters. We see that for about half the clusters, there is a majority language family, most often Niger-Congo. We also observe distinctive clusters devoted to Austronesian and Quechuan languages. The largest two clusters are rather indistinct, without any single language family achieving more than 24% of the total.

8 Conclusion

In this paper, we presented a successful solution to one aspect of the decipherment task: the prediction of consonants and vowels for an unknown language and alphabet. Adopting a classical Bayesian perspective, we develop a model that performs posterior inference over hundreds of languages, leveraging knowledge of known languages to uncover general linguistic patterns of typologically coherent language clusters. Using this model, we automatically distinguish between consonant and vowel characters with nearly 99% accuracy across 503 languages. We further experimented on a three-way classification task involving nasal characters, achieving nearly 90% accuracy.

Future work will take us in several new directions: first, we would like to move beyond the assumption of an alphabetic writing system so that we can apply our method to undeciphered syllabic scripts such as Linear A. We would also like to extend our methods to achieve finer-grained resolution of phonetic properties beyond nasals, consonants, and vowels.

Acknowledgments

The authors thank the reviewers and acknowledge support by the NSF (grant IIS-1116676) and a research gift from Google. Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the NSF.

References

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of the ACL*, pages 1288–1297. Association for Computational Linguistics.

Alexandre Bouchard-Côté, David Hall, Thomas L Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2011. A Bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of EMNLP*, pages 638–647. Association for Computational Linguistics.

Shay B Cohen and Noah A Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of NAACL*, pages 74–82. Association for Computational Linguistics.

Shay B Cohen, Dipanjan Das, and Noah A Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of EMNLP*, pages 50–61. Association for Computational Linguistics.

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

Young-Bum Kim and Benjamin Snyder. 2012. Universal grapheme-to-phoneme prediction over latin alphabets. In *Proceedings of EMNLP*, pages 332–343, Jeju Island, South Korea, July. Association for Computational Linguistics.

Young-Bum Kim, João V Graça, and Benjamin Snyder. 2011. Universal morphological analysis using structured nearest neighbor prediction. In *Proceedings of EMNLP*, pages 322–332. Association for Computational Linguistics.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of COLING/ACL*, pages 499–506. Association for Computational Linguistics.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2010. Simple type-level unsupervised POS tagging. In *Proceedings of EMNLP*, pages 853–861. Association for Computational Linguistics.

Percy Liang, Michael I Jordan, and Dan Klein. 2010. Type-based MCMC. In *Proceedings of NAACL*, pages 573–581. Association for Computational Linguistics.

David JC MacKay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36(1):341–385.

Radford M Neal. 2003. Slice sampling. *Annals of statistics*, 31:705–741.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the ACL*, pages 1048–1057. Association for Computational Linguistics.