Improved Reconstruction of Protolanguage Word Forms

Alexandre Bouchard-Côté* **Thomas L. Griffiths†** **Dan Klein***
*Computer Science Division †Department of Psychology
University of California at Berkeley
Berkeley, CA 94720

Abstract

We present an unsupervised approach to reconstructing ancient word forms. The present work addresses three limitations of previous work. First, previous work focused on *faithfulness* features, which model changes between successive languages. We add *markedness* features, which model well-formedness within each language. Second, we introduce universal features, which support generalizations across languages. Finally, we increase the number of languages to which these methods can be applied by an order of magnitude by using improved inference methods. Experiments on the reconstruction of Proto-Oceanic, Proto-Malayo-Javanic, and Classical Latin show substantial reductions in error rate, giving the best results to date.

1 Introduction

A central problem in diachronic linguistics is the reconstruction of ancient languages from their modern descendants (Campbell, 1998). Here, we consider the problem of reconstructing phonological forms, given a known linguistic phylogeny and known cognate groups. For example, Figure 1 (a) shows a collection of word forms in several Oceanic languages, all meaning *to cry*. The ancestral form in this case has been presumed to be /taŋis/ in Blust (1993). We are interested in models which take as input many such word tuples, each representing a cognate group, along with a language tree, and induce word forms for hidden ancestral languages.

The traditional approach to this problem has been the *comparative method*, in which reconstructions are done manually using assumptions about the relative probability of different kinds of sound change (Hock, 1986). There has been work attempting to automate part (Durham and Rogers, 1969; Eastlack, 1977; Lowe and Mazaudon, 1994; Covington, 1998; Kondrak, 2002) or all of the process (Oakes, 2000; Bouchard-Côté et al., 2008). However, previous automated methods have been unable to leverage three important ideas a linguist would employ. We address these omissions here, resulting in a more powerful method for automatically reconstructing ancient protolanguages.

First, linguists triangulate reconstructions from many languages, while past work has been limited to small numbers of languages. For example, Oakes (2000) used four languages to reconstruct Proto-Malayo-Javanic (PMJ) and Bouchard-Côté et al. (2008) used two languages to reconstruct Classical Latin (La). We revisit these small datasets and show that our method significantly outperforms these previous systems. However, we also show that our method can be applied to a much larger data set (Greenhill et al., 2008), reconstructing Proto-Oceanic (POc) from 64 modern languages. In addition, performance *improves* with more languages, which was not the case for previous methods.

Second, linguists exploit knowledge of phonological universals. For example, small changes in vowel height or consonant place are more likely than large changes, and much more likely than change to arbitrarily different phonemes. In a statistical system, one could imagine either manually encoding or automatically inferring such preferences. We show that both strategies are effective.

Finally, linguists consider not only how languages change, but also how they are internally consistent. Past models described how sounds do (or, more often, do not) change between nodes in the tree. To borrow broad terminology from the Optimality Theory literature (Prince and Smolensky, 1993), such models incorporated *faithfulness* features, capturing the ways in which successive forms remained similar to one another. However, each language has certain regular phonotactic patterns which con-

strain these changes. We encode such patterns using *markedness* features, characterizing the internal phonotactic structure of each language. Faithfulness and markedness play roles analogous to the channel and language models of a noisy-channel system. We show that markedness features improve reconstruction, and can be used efficiently.

2 Related work

Our focus in this section is on describing the properties of the two previous systems for reconstructing ancient word forms to which we compare our method. Citations for other related work, such as similar approaches to using faithfulness and markedness features, appear in the body of the paper.

In Oakes (2000), the word forms in a given protolanguage are reconstructed using a Viterbi multi-alignment between a small number of its descendant languages. The alignment is computed using hand-set parameters. Deterministic rules characterizing changes between pairs of observed languages are extracted from the alignment when their frequency is higher than a threshold, and a proto-phoneme inventory is built using linguistically motivated rules and parsimony. A reconstruction of each observed word is first proposed independently for each language. If at least two reconstructions agree, a majority vote is taken, otherwise no reconstruction is proposed. This approach has several limitations. First, it is not tractable for larger trees, since the time complexity of their multi-alignment algorithm grows exponentially in the number of languages. Second, deterministic rules, while elegant in theory, are not robust to noise: even in experiments with only four daughter languages, a large fraction of the words could not be reconstructed.

In Bouchard-Côté et al. (2008), a stochastic model of sound change is used and reconstructions are inferred by performing probabilistic inference over an evolutionary tree expressing the relationships between languages. The model does not support generalizations across languages, and has no way to capture phonotactic regularities within languages. As a consequence, the resulting method does not scale to large phylogenies. The work we present here addresses both of these issues, with a richer model and faster inference allowing improved reconstruction and increased scale.

3 Model

We start this section by introducing some notation. Let τ be a tree of languages, such as the examples in Figure 3 (c-e). In such a tree, the modern languages, whose word forms will be observed, are the leaves of τ. All internal nodes, particularly the root, are languages whose word forms are not observed. Let L denote all languages, modern and otherwise. All word forms are assumed to be strings Σ^* in the International Phonological Alphabet (IPA).[1]

We assume that word forms evolve along the branches of the tree τ. However, it is not the case that each cognate set exists in each modern language. Formally, we assume there to be a known list of C cognate sets. For each $c \in \{1, \ldots, C\}$ let $L(c)$ denote the subset of modern languages that have a word form in the c-th cognate set. For each set $c \in \{1, \ldots, C\}$ and each language $\ell \in L(c)$, we denote the modern word form by $w_{c\ell}$. For cognate set c, only the minimal subtree $\tau(c)$ containing $L(c)$ and the root is relevant to the reconstruction inference problem for that set.

From a high-level perspective, the generative process is quite simple. Let c be the index of the current cognate set, with topology $\tau(c)$. First, a word is generated for the root of $\tau(c)$ using an (initially unknown) root language model (distribution over strings). The other nodes of the tree are drawn incrementally as follows: for each edge $\ell \rightarrow \ell'$ in $\tau(c)$ use a branch-specific distribution over changes in strings to generate the word at node ℓ'.

In the remainder of this section, we clarify the exact form of the conditional distributions over string changes, the distribution over strings at the root, and the parameterization of this process.

3.1 Markedness and Faithfulness

In Optimality Theory (OT) (Prince and Smolensky, 1993), two types of constraints influence the selection of a realized output given an input form: *faithfulness* and *markedness* constraints. Faithfulness en-

[1] The choice of a phonemic representation is motivated by the fact that most of the data available comes in this form. Diacritics are available in a smaller number of languages and may vary across dialects, so we discarded them in this work.

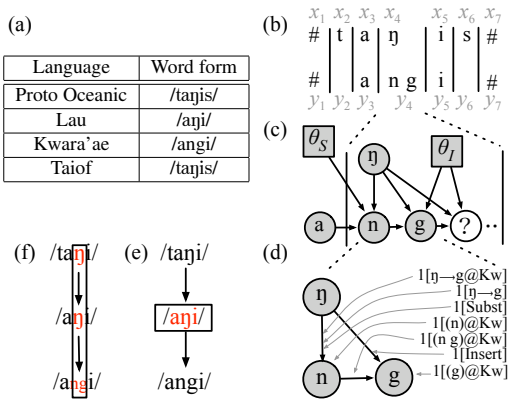

Figure 1: (a) A cognate set from the Austronesian dataset. All word forms mean *to cry*. (b-d) The mutation model used in this paper. (b) The mutation of POc /taŋis/ to Kw. /angi/. (c) Graphical model depicting the dependencies among variables in one step of the mutation Markov chain. (d) Active features for one step in this process. (e-f) Comparison of two inference procedures on trees: Single sequence resampling (e) draws one sequence at a time, conditioned on its parent and children, while ancestry resampling (f) draws an aligned slice from all words simultaneously. In large trees, the latter is more efficient than the former.

courages similarity between the input and output while markedness favors well-formed output.

Viewed from this perspective, previous computational approaches to reconstruction are based almost exclusively on faithfulness, expressed through a mutation model. Only the words in the language at the root of the tree, if any, are explicitly encouraged to be well-formed. In contrast, we incorporate constraints on markedness for each language with both general and branch-specific constraints on faithfulness. This is done using a *lexicalized stochastic string transducer* (Varadarajan et al., 2008).

We now make precise the conditional distributions over pairs of evolving strings, referring to Figure 1 (b-d). Consider a language ℓ' evolving to ℓ for cognate set c. Assume we have a word form $x = w_{c\ell'}$. The generative process for producing $y = w_{c\ell}$ works as follows. First, we consider x to be composed of characters $x_1 x_2 \ldots x_n$, with the first and last being a special boundary symbol $x_1 = \# \in \Sigma$ which is never deleted, mutated, or created. The process generates $y = y_1 y_2 \ldots y_n$ in n chunks $y_i \in \Sigma^*, i \in \{1, \ldots, n\}$, one for each x_i. The y_i's may be a single character, multiple characters, or even empty. In the example shown, all three of these cases occur.

To generate y_i, we define a *mutation Markov chain* that incrementally adds zero or more characters to an initially empty y_i. First, we decide whether the current phoneme in the top word $t = x_i$ will be deleted, in which case $y_i = \epsilon$ as in the example of /s/ being deleted. If t is not deleted, we chose a single substitution character in the bottom word. This is the case both when /a/ is unchanged and when /ŋ/ substitutes to /n/. We write $\mathscr{S} = \Sigma \cup \{\zeta\}$ for this set of outcomes, where ζ is the special outcome indicating deletion. Importantly, the probabilities of this multinomial can depend on both the previous character generated so far (i.e. the rightmost character p of y_{i-1}) and the current character in the previous generation string (t). As we will see shortly, this allows modelling markedness and faithfulness at every branch, jointly. This multinomial decision acts as the initial distribution of the mutation Markov chain.

We consider insertions only if a deletion was not selected in the first step. Here, we draw from a multinomial over $\mathscr{S}$, where this time the special outcome ζ corresponds to stopping insertions, and the other elements of $\mathscr{S}$ correspond to symbols that are appended to y_i. In this case, the conditioning environment is $t = x_i$ and the current rightmost symbol p in y_i. Insertions continue until ζ is selected. In the example, we follow the substitution of /ŋ/ to /n/ with an insertion of /g/, followed by a decision to stop that y_i. We will use $\theta_{S,t,p,\ell}$ and $\theta_{I,t,p,\ell}$ to denote the probabilities over the substitution and insertion decisions in the current branch $\ell' \rightarrow \ell$.

A similar process generates the word at the root ℓ of a tree, treating this word as a single string y_1 generated from a dummy ancestor $t = x_1$. In this case, only the insertion probabilities matter, and we separately parameterize these probabilities with $\theta_{R,t,p,\ell}$. There is no actual dependence on t at the root, but this formulation allows us to unify the parameterization, with each $\theta_{\omega,t,p,\ell} \in \mathbb{R}^{|\Sigma|+1}$ where $\omega \in \{R, S, I\}$.

3.2 Parameterization

Instead of directly estimating the transition probabilities of the mutation Markov chain (as the parameters of a collection of multinomial distributions) we

express them as the output of a log-linear model. We used the following feature templates:

OPERATION identifies whether an operation in the mutation Markov chain is an insertion, a deletion, a substitution, a self-substitution (i.e. of the form $x \rightarrow y, x = y$), or the end of an insertion event. Examples in Figure 1 (d): $\mathbf{1}[\text{Subst}]$ and $\mathbf{1}[\text{Insert}]$.

MARKEDNESS consists of language-specific n-gram indicator functions for all symbols in Σ. Only unigram and bigram features are used for computational reasons, but we show in Section 5 that this already captures important constraints. Examples in Figure 1 (d): the bigram indicator $\mathbf{1}[(\text{n g})@\text{Kw}]$ (Kw stands for Kwara'ae, a language of the Solomon Islands), the unigram indicators $\mathbf{1}[(\text{n})@\text{Kw}]$ and $\mathbf{1}[(\text{g})@\text{Kw}]$.

FAITHFULNESS consists of indicators for mutation events of the form $\mathbf{1}[x \rightarrow y]$, where $x \in \Sigma$, $y \in \mathscr{S}$. Examples: $\mathbf{1}[\text{ŋ} \rightarrow \text{n}]$, $\mathbf{1}[\text{ŋ} \rightarrow \text{n}@\text{Kw}]$.

Feature templates similar to these can be found for instance in Dreyer et al. (2008) and Chen (2003), in the context of string-to-string transduction. Note also the connection with stochastic OT (Goldwater and Johnson, 2003; Wilson, 2006), where a log-linear model mediates markedness and faithfulness of the production of an output form from an underlying input form.

3.3 Parameter sharing

Data sparsity is a significant challenge in protolanguage reconstruction. While the experiments we present here use an order of magnitude more languages than previous computational approaches, the increase in observed data also brings with it additional unknowns in the form of intermediate protolanguages. Since there is one set of parameters for each language, adding more data is not sufficient for increasing the quality of the reconstruction: we show in Section 5.2 that adding extra languages can actually hurt reconstruction using previous methods. It is therefore important to share parameters across different branches in the tree in order to benefit from having observations from more languages.

As an example of useful parameter sharing, consider the faithfulness features $\mathbf{1}[\text{/p/} \rightarrow \text{/b/}]$ and $\mathbf{1}[\text{/p/} \rightarrow \text{/r/}]$, which are indicator functions for the appearance of two substitutions for /p/. We would like the model to learn that the former event (a simple voicing change) should be preferred over the latter. In Bouchard-Côté et al. (2008), this has to be learned for each branch in the tree. The difficulty is that not all branches will have enough information to learn this preference, meaning that we need to define the model in such a way that it can generalize across languages.

We used the following technique to address this problem: we augment the sufficient statistics of Bouchard-Côté et al. (2008) to include the current language (or language at the bottom of the current branch) and use a single, global weight vector instead of a set of branch-specific weights. Generalization across branches is then achieved by using features that *ignore* ℓ, while branch-specific features depend on ℓ.

For instance, in Figure 1 (d), $\mathbf{1}[\text{ŋ} \rightarrow \text{n}]$ is an example of a universal (global) feature shared across all branches while $\mathbf{1}[\text{ŋ} \rightarrow \text{n}@\text{Kw}]$ is branch-specific. Similarly, all of the features in OPERATION, MARKEDNESS and FAITHFULNESS have universal and branch-specific versions.

3.4 Objective function

Concretely, the transition probabilities of the mutation and root generation are given by:

$$\theta_{\omega,t,p,\ell}(\xi) = \frac{\exp\{\langle \lambda, f(\omega,t,p,\ell,\xi)\rangle\}}{Z(\omega,t,p,\ell,\lambda)} \times \mu(\omega,t,\xi),$$

where $\xi \in \mathscr{S}$, $f : \{S,I,R\} \times \Sigma \times \Sigma \times L \times \mathscr{S} \rightarrow \mathbb{R}^k$ is the sufficient statistics or feature function, $\langle \cdot, \cdot \rangle$ denotes inner product and $\lambda \in \mathbb{R}^k$ is a weight vector. Here, k is the dimensionality of the feature space of the log-linear model. In the terminology of exponential families, Z and μ are the normalization function and reference measure respectively:

$$Z(\omega,t,p,\ell,\lambda) = \sum_{\xi' \in \mathscr{S}} \exp\{\langle \lambda, f(\omega,t,p,\ell,\xi')\rangle\}$$

$$\mu(\omega,t,\xi) = \begin{cases} 0 \text{ if } \omega = S, t = \#, \xi \neq \# \\ 0 \text{ if } \omega = R, \xi = \zeta \\ 0 \text{ if } \omega \neq R, \xi = \# \\ 1 \text{ o.w.} \end{cases}$$

Here, μ is used to handle boundary conditions.

We will also need the following notation: let $\mathbb{P}_\lambda(\cdot), \mathbb{P}_\lambda(\cdot|\cdot)$ denote the root and branch probability models described in Section 3.1 (with transition probabilities given by the above log-linear model), $I(c)$, the set of internal (non-leaf) nodes in $\tau(c)$, $\text{pa}(\ell)$, the parent of language ℓ, $\text{r}(c)$, the root of $\tau(c)$

and $W(c) = (\Sigma^*)^{|I(c)|}$. We can summarize our objective function as follows:

$$\sum_{c=1}^{C} \log \sum_{\vec{w} \in W(c)} \mathbb{P}_\lambda(w_{c,\mathrm{r}(c)}) \prod_{\ell \in I(c)} \mathbb{P}_\lambda(w_{c,\ell} | w_{c,\mathrm{pa}(\ell)}) - \frac{||\lambda||_2^2}{2\sigma^2}$$

The second term is a standard L^2 regularization penalty (we used $\sigma^2 = 1$).

4 Learning algorithm

Learning is done using a Monte Carlo variant of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The M step is convex and computed using L-BFGS (Liu et al., 1989); but the E step is intractable (Lunter et al., 2003), so we used a Markov chain Monte Carlo (MCMC) approximation (Tierney, 1994). At E step $t = 1, 2, \ldots$, we simulated the chain for $O(t)$ iterations; this regime is necessary for convergence (Jank, 2005).

In the E step, the inference problem is to compute an expectation under the posterior over strings in a protolanguage given observed word forms at the leaves of the tree. The typical approach in biology or historical linguistics (Holmes and Bruno, 2001; Bouchard-Côté et al., 2008) is to use Gibbs sampling, where the entire string at a single node in the tree is sampled, conditioned on its parent and children. This sampling domain is shown in Figure 1 (e), where the middle word is completely resampled but adjacent words are fixed. We will call this method Single Sequence Resampling (SSR). While conceptually simple, this approach suffers from problems in large trees (Holmes and Bruno, 2001). Consequently, we use a different MCMC procedure, called Ancestry Resampling (AR) that alleviates the mixing problems (Figure 1 (f)). This method was originally introduced for biological applications (Bouchard-Côté et al., 2009), but commonalities between the biological and linguistic cases make it possible to use it in our model.

Concretely, the problem with SSR arises when the tree under consideration is large or unbalanced. In this case, it can take a long time for information from the observed languages to propagate to the root of the tree. Indeed, samples at the root will initially be *independent* of the observations. AR addresses this problem by resampling one thin vertical slice of all sequences at a time, called an ancestry. For the precise definition, see Bouchard-Côté et al. (2009). Slices condition on observed data, avoiding the problems mentioned above, and can propagate information rapidly across the tree.

5 Experiments

We performed a comprehensive set of experiments to test the new method for reconstruction outlined above. In Section 5.1, we analyze in isolation the effects of varying the set of features, the number of observed languages, the topology, and the number of iterations of EM. In Section 5.2 we compare performance to an oracle and to three other systems.

Evaluation of all methods was done by computing the Levenshtein distance (Levenshtein, 1966) between the reconstruction produced by each method and the reconstruction produced by linguists. We averaged this distance across reconstructed words to report a single number for each method. We show in Table 2 the average word length in each corpus; note that the Latin average is much larger, giving an explanation to the higher errors in the Romance dataset. The statistical significance of all performance differences are assessed using a paired t-test with significance level of 0.05.

5.1 Evaluating system performance

We used the Austronesian Basic Vocabulary Database (Greenhill et al., 2008) as the basis for a series of experiments used to evaluate the performance of our system and the factors relevant to its success. The database includes partial cognacy judgments and IPA transcriptions, as well as a few reconstructed protolanguages. A reconstruction of Proto-Oceanic (POc) originally developed by Blust (1993) using the comparative method was the basis for evaluation.

We used the cognate information provided in the database, automatically constructing a global tree[2] and set of subtrees from the cognate set indicator matrix $M(\ell, c) = \mathbf{1}[\ell \in L(c)], c \in \{1, \ldots, C\}, \ell \in L$. For constructing the global tree, we used the implementation of neighbor joining in the Phylip package (Felsenstein, 1989). We used a distance based on cognates overlap, $d_c(\ell_1, \ell_2) = \sum_{c=1}^{C} M(\ell_1, c) M(\ell_2, c)$. We bootstrapped 1000

[2]The dataset included a tree, but it was out of date as of November 2008 (Greenhill et al., 2008).

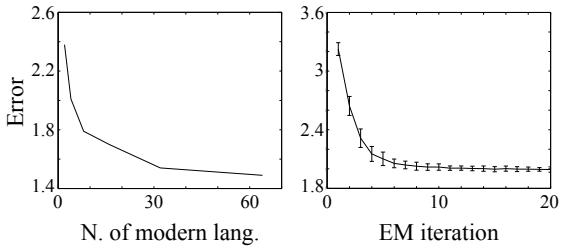

Figure 2: Left: Mean distance to the target reconstruction of POc as a function of the number of modern languages used by the inference procedure. Right: Mean distance and confidence intervals as a function of the EM iteration, averaged over 20 random seeds and ran on 4 languages.

Condition	Edit dist.
Unsupervised full system	1.87
-FAITHFULNESS	2.02
-MARKEDNESS	2.18
-Sharing	1.99
-Topology	2.06
Semi-supervised system	1.75

Table 1: Effects of ablation of various aspects of our unsupervised system on mean edit distance to POc. -Sharing corresponds to the restriction to the subset of the features in OPERATION, FAITHFULNESS and MARKEDNESS that are branch-specific, -Topology corresponds to using a flat topology where the only edges in the tree connect modern languages to POc. The semi-supervised system is described in the text. All differences (compared to the unsupervised full system) are statistically significant.

samples and formed an accurate (90%) consensus tree. The tree obtained is not binary, but the AR inference algorithm scales linearly in the branching factor of the tree (in contrast, SSR scales exponentially (Lunter et al., 2003)).

The first claim we verified experimentally is that having more observed languages aids reconstruction of protolanguages. To test this hypothesis we added observed modern languages in increasing order of distance d_c to the target reconstruction of POc so that the languages that are most useful for POc reconstruction are added first. This prevents the effects of adding a close language after several distant ones being confused with an improvement produced by increasing the number of languages.

The results are reported in Figure 2 (a). They confirm that large-scale inference is desirable for automatic protolanguage reconstruction: reconstruction improved statistically significantly with each increase except from 32 to 64 languages, where the average edit distance improvement was 0.05.

We then conducted a number of experiments intended to assess the robustness of the system, and to identify the contribution made by different factors it incorporates. First, we ran the system with 20 different random seeds to assess the stability of the solutions found. In each case, learning was stable and accuracy improved during training. See Figure 2 (b).

Next, we found that all of the following ablations significantly hurt reconstruction: using a flat tree (in which all languages are equidistant from the reconstructed root and from each other) instead of the consensus tree, dropping the markedness features, dropping the faithfulness features, and disabling sharing across branches. The results of these experiments are shown in Table 1.

For comparison, we also included in the same table the performance of a semi-supervised system trained by K-fold validation. The system was ran $K = 5$ times, with $1 - K^{-1}$ of the POc words given to the system as observations in the graphical model for each run. It is semi-supervised in the sense that gold reconstruction for many internal nodes are not available in the dataset (for example the common ancestor of Kwara'ae (Kw.) and Lau in Figure 3 (b)), so they are still not filled.[3]

Figure 3 (b) shows the results of a concrete run over 32 languages, zooming in to a pair of the Solomonic languages and the cognate set from Figure 1 (a). In the example shown, the reconstruction is as good as the ORACLE (described in Section 5.2), though off by one character (the final /s/ is not present in any of the 32 inputs and therefore is not reconstructed). In (a), diagrams show, for both the global and the local (Kwara'ae) features, the expectations of each substitution superimposed on an IPA sound chart, as well as a list of the top changes. Darker lines indicate higher counts. This run did not use natural class constraints, but it can

[3]We also tried a fully supervised system where a flat topology is used so that all of these latent internal nodes are avoided; but it did not perform as well—this is consistent with the -Topology experiment of Table 1.

be seen that linguistically plausible substitutions are learned. The global features prefer a range of voicing changes, manner changes, adjacent vowel motion, and so on, including mutations like /s/ to /h/ which are common but poorly represented in a naive attribute-based natural class scheme. On the other hand, the features local to the language Kwara'ae pick out the subset of these changes which are active in that branch, such as /s/→/t/ fortition.

5.2 Comparisons against other methods

The first two competing methods, PRAGUE and BCLKG, are described in Oakes (2000) and Bouchard-Côté et al. (2008) respectively and summarized in Section 1. Neither approach scales well to large datasets. In the first case, the bottleneck is the complexity of computing multi-alignments without guide trees and the vanishing probability that independent reconstructions agree. In the second case, the problem comes from the unregularized proliferation of parameters and slow mixing of the inference algorithm. For this reason, we built a third baseline that scales well in large datasets.

This third baseline, CENTROID, computes the centroid of the observed word forms in Levenshtein distance. Let $L(x, y)$ denote the Levenshtein distance between word forms x and y. Ideally, we would like the baseline to return $\text{argmin}_{x \in \Sigma^*} \sum_{y \in O} L(x, y)$, where $O = \{y_1, \ldots, y_{|O|}\}$ is the set of observed word forms. Note that the optimum is not changed if we restrict the minimization to be taken on $x \in \Sigma(O)^*$ such that $m \leq |x| \leq M$ where $m = \min_i |y_i|, M = \max_i |y_i|$ and $\Sigma(O)$ is the set of characters occurring in O. Even with this restriction, this optimization is intractable. As an approximation, we considered only strings built by at most k contiguous substrings taken from the word forms in O. If $k = 1$, then it is equivalent to taking the min over $x \in O$. At the other end of the spectrum, if $k = M$, it is exact. This scheme is exponential in k, but since words are relatively short, we found that $k = 2$ often finds the same solution as higher values of k. The difference was in all the cases not statistically significant, so we report the approximation $k = 2$ in what follows.

We also compared against an oracle, denoted ORACLE, which returns $\text{argmin}_{y \in O} L(y, x^*)$, where x^* is the target reconstruction. We will denote it by ORACLE. This is superior to picking a single closest language to be used for all word forms, but it is possible for systems to perform better than the oracle since it has to return one of the observed word forms.

Comparison	CENTROID	PRAGUE	BCLKG
Protolanguage	POc	PMJ	La
Heldout (prop.)	243 (1.0)	79 (1.0)	293 (0.5)
Modern languages	70	4	2
Cognate sets	1321	179	583
Observed words	10783	470	1463
Mean word length	4.5	5.0	7.4

Table 2: Experimental setup: number of held-out protoword from (absolute and relative), of modern languages, cognate sets and total observed words. The split for BCLKG is the same as in Bouchard-Côté et al. (2008).

We performed the comparison against Oakes (2000) and Bouchard-Côté et al. (2008) on the same dataset and experimental conditions as those used in the respective papers (see Table 2). Note that the setup of Bouchard-Côté et al. (2008) provides supervision (half of the Latin word forms are provided); all of the other comparisons are performed in a completely unsupervised manner.

The PMJ dataset was compiled by Nothofer (1975), who also reconstructed the corresponding protolanguage. Since PRAGUE is not guaranteed to return a reconstruction for each cognate set, only 55 word forms could be directly compared to our system. We restricted comparison to this subset of the data. This favors PRAGUE since the system only proposes a reconstruction when it is certain. Still, our system outperformed PRAGUE, with an average distance of 1.60 compared to 2.02 for PRAGUE. The difference is marginally significant, $p = 0.06$, partly due to the small number of word forms involved.

We also exceeded the performance of BCLKG on the Romance dataset. Our system's reconstruction had an edit distance of 3.02 to the truth against 3.10 for BCLKG. However, this difference was not significant ($p = 0.15$). We think this is because of the high level of noise in the data (the Romance dataset is the only dataset we consider that was automatically constructed rather than curated by linguists). A second factor contributing to this small difference may be that the the experimental setup of BCLKG used very few languages, while the performance of our system improves markedly with more languages.

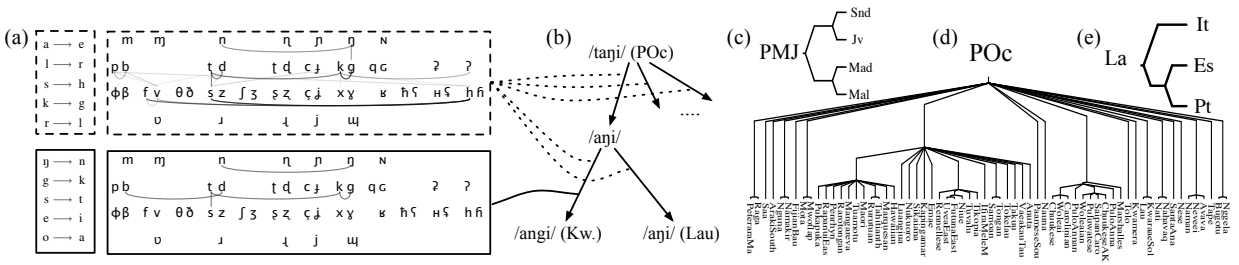

Figure 3: (a) A visualization of two learned faithfulness parameters: on the top, from the universal features, on the bottom, for one particular branch. Each pair of phonemes have a link with grayscale value proportional to the expectation of a transition between them. The five strongest links are also included at the right. (b) A sample taken from our POc experiments (see text). (c-e) Phylogenetic trees for three language families: Proto-Malayo-Javanic, Austronesian and Romance.

We conducted another experiment to verify this by running both systems in larger trees. Because the Romance dataset had only three modern languages transcribed in IPA, we used the Austronesian dataset to perform the test. The results were all significant in this setup: while our method went from an edit distance of 2.01 to 1.79 in the 4-to-8 languages experiment described in Section 5.1, BCLKG went from 3.30 to 3.38. This suggests that more languages can actually *hurt* systems that do not support parameter sharing.

Since we have shown evidence that PRAGUE and BCLKG do not scale well to large datasets, we also compared against ORACLE and CENTROID in a large-scale setting. Specifically, we compare to the experimental setup on 64 modern languages used to reconstruct POc described before. Encouragingly, while the system's average distance (1.49) does not attain that of the ORACLE (1.13), we significantly outperform the CENTROID baseline (1.79).

5.3 Incorporating prior linguistic knowledge

The model also supports the addition of prior linguistic knowledge. This takes the form of feature templates with more internal structure. We performed experiments with an additional feature template:

STRUCT-FAITHFULNESS is a structured version of FAITHFULNESS, replacing x and y with their natural classes $N_\beta(x)$ and $N_\beta(y)$ where β indexes types of classes, ranging over {manner, place, phonation, isOral, isCentral, height, backness, roundedness}. This feature set is reminiscent of the featurized representation of Kondrak (2000).

We compared the performance of the system with and without STRUCT-FAITHFULNESS to check if the algorithm can recover the structure of natural classes in an unsupervised fashion. We found that with 2 or 4 observed languages, FAITHFULNESS underperformed STRUCT-FAITHFULNESS, but for larger trees, the difference was not significant. FAITHFULNESS even slightly outperformed its structured cousin with 16 observed languages.

6 Conclusion

By enriching our model to include important features like markedness, and by scaling up to much larger data sets than were previously possible, we obtained substantial improvements in reconstruction quality, giving the best results on past data sets. While many more complex phenomena are still unmodeled, from reduplication to borrowing to chained sound shifts, the current approach significantly increases the power, accuracy, and efficiency of automatic reconstruction.

Acknowledgments

We would like to thank Anna Rafferty and our reviewers for their comments. This work was supported by a NSERC fellowship to the first author and NSF grant number BCS-0631518 to the second author.

References

R. Blust. 1993. Central and central-Eastern Malayo-Polynesian. *Oceanic Linguistics*, 32:241–293.

A. Bouchard-Côté, P. Liang, D. Klein, and T. L. Griffiths. 2008. A probabilistic approach to language change. In *Advances in Neural Information Processing Systems 20*.

A. Bouchard-Côté, M. I. Jordan, and D. Klein. 2009. Efficient inference in phylogenetic InDel trees. In *Advances in Neural Information Processing Systems 21*.

L. Campbell. 1998. *Historical Linguistics*. The MIT Press.

S. F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proceedings of Eurospeech*.

M. A. Covington. 1998. Alignment of multiple languages for historical comparison. In *Proceedings of ACL 1998*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

M. Dreyer, J. R. Smith, and J. Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of EMNLP 2008*.

S. P. Durham and D. E. Rogers. 1969. An application of computer programming to the reconstruction of a proto-language. In *Proceedings of the 1969 conference on Computational linguistics*.

C. L. Eastlack. 1977. Iberochange: A program to simulate systematic sound change in Ibero-Romance. *Computers and the Humanities*.

J. Felsenstein. 1989. PHYLIP - PHYLogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166.

S. Goldwater and M. Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Workshop on Variation within Optimality Theory*.

S. J. Greenhill, R. Blust, and R. D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283.

H. H. Hock. 1986. *Principles of Historical Linguistics*. Walter de Gruyter.

I. Holmes and W. J. Bruno. 2001. Evolutionary HMM: a Bayesian approach to multiple alignment. *Bioinformatics*, 17:803–820.

W. Jank. 2005. Stochastic variants of EM: Monte Carlo, quasi-Monte Carlo and more. In *Proceedings of the American Statistical Association*.

G. Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000*.

G. Kondrak. 2002. *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, February.

D. C. Liu, J. Nocedal, and C. Dong. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

J. B. Lowe and M. Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Comput. Linguist.*, 20(3):381–417.

G. A. Lunter, I. Miklós, Y. S. Song, and J. Hein. 2003. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *Journal of Computational Biology*, 10:869–889.

B. Nothofer. 1975. *The reconstruction of Proto-Malayo-Javanic*. M. Nijhoff.

M. P. Oakes. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–244.

A. Prince and P. Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science.

L. Tierney. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.

A. Varadarajan, R. K. Bradley, and I. H. Holmes. 2008. Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology*, 9:R147.

C. Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30.5:945–982.