Large-Scale Cognate Recovery

David Hall and **Dan Klein**
Computer Science Division
University of California at Berkeley
{dlwh,klein}@cs.berkeley.edu

Abstract

We present a system for the large scale induction of cognate groups. Our model explains the evolution of cognates as a sequence of mutations and innovations along a phylogeny. On the task of identifying cognates from over 21,000 words in 218 different languages from the Oceanic language family, our model achieves a cluster purity score over 91%, while maintaining pairwise recall over 62%.

1 Introduction

The critical first step in the reconstruction of an ancient language is the recovery of related *cognate* words in its descendants. Unfortunately, this process has largely been a manual, linguistically-intensive undertaking for any sizable number of descendant languages. The traditional approach used by linguists—the comparative method—iterates between positing putative cognates and then identifying regular sound laws that explain correspondences between those words (Bloomfield, 1938).

Successful computational approaches have been developed for large-scale reconstruction of phylogenies (Ringe et al., 2002; Daumé III and Campbell, 2007; Daumé III, 2009; Nerbonne, 2010) and ancestral word forms of known cognate sets (Oakes, 2000; Bouchard-Côté et al., 2007; Bouchard-Côté et al., 2009), enabling linguists to explore deep historical relationships in an automated fashion. However, computational approaches thus far have not been able to offer the same kind of scale for identifying cognates. Previous work in cognate identification has largely focused on identifying cognates in pairs of languages (Mann and Yarowsky, 2001; Lowe and Mazaudon, 1994; Oakes, 2000; Kondrak, 2001; Mulloni, 2007), with a few recent exceptions that can find sets in a handful of languages (Bergsma and Kondrak, 2007; Hall and Klein, 2010).

While it may seem surprising that cognate detection has not successfully scaled to large numbers of languages, the task poses challenges not seen in reconstruction and phylogeny inference. For instance, morphological innovations and irregular sound changes can completely obscure relationships between words in different languages. However, in the case of reconstruction, an unexplainable word is simply that: one can still correctly reconstruct its ancestor using words from related languages.

In this paper, we present a system that uses two generative models for large-scale cognate identification. Both models describe the evolution of words along a phylogeny according to automatically learned sound laws in the form of parametric edit distances. The first is an adaptation of the generative model of Hall and Klein (2010), and the other is a new generative model called PARSIM with connections to parsimony methods in computational biology (Cavalli-Sforza and Edwards, 1965; Fitch, 1971). Our model supports simple, tractable inference via message passing, at the expense of being unable to model some cognacy relationships. To help correct this deficiency, we also describe an agglomerative inference procedure for the model of Hall and Klein (2010). By using the output of our system as input to this system, we can find cognate groups that PARSIM alone cannot recover.

We apply these models to identifying cognate groups from two language families using the Austronesian Basic Vocabulary Database (Greenhill et al., 2008), a catalog of words from about 40% of the Austronesian languages. We focus on data from two subfamilies of Austronesian: Formosan and

Oceanic. The datasets are by far the largest on which automated cognate recovery has ever been attempted, with 18 and 271 languages respectively. On the larger Oceanic data, our model can achieve cluster purity scores of 91.8%, while maintaining pairwise recall of 62.1%. We also analyze the mistakes of our system, where we find that some of the erroneous cognate groups our system finds may not be errors at all. Instead, they may be previously unknown cognacy relationships that were not annotated in the data.

2 Background

Before we present our model, we first describe basic facts of the Austronesian language family, along with a description of the Austronesian Basic Vocabulary Database, which forms the dataset that we use for our experiments. For far more detailed coverage of the Austronesian languages, we direct the interested reader to Blust (2009)'s comprehensive monograph.

2.1 The Austronesian Language Family

The Austronesian language family is one of the largest in the world, comprising about one-fifth of the world's languages. Geographically, it stretches from its homeland on Formosa (Taiwan) to Madagascar in the west, and as far as Hawai'i and (at one point) the Easter Islands to the east. Until the advent of European colonialism spread Indo-European languages to every continent, Austronesian was the most widespread of all language families.

Linguistically, the language family is as diverse as it is large, but a few regularities hold. From a phonological perspective, two features stand out. First, the phoneme inventories of these languages are typically small. For example, it is well-known that Hawaiian has only 13 phonemes. Moreover, the phonotactics of these languages are often restrictive. Sticking with the same example, Hawaiian only allows (C)V syllables: consonants clusters are forbidden, and no syllable may end with a consonant.

2.2 The Austronesian Basic Vocabulary Database

The Austronesian Basic Vocabulary Database (ABVD) (Greenhill et al., 2008) is an ambitious, ongoing effort to catalog the lexicons and basic facts about all of the languages in the Austronesian language family. It also contains manual reconstructions for select ancestor languages produced by linguists.

The sample we use—from Bouchard-Côté et al. (2009)—contains about 50,000 words across 471 languages spanning all the major divisions of Austronesian. These words are grouped into cognate groups and arranged by gloss. For instance, there are 37 distinct cognate groups for the gloss "tail." One of these groups includes the words */ekor/*, */ingko/*, */iŋkot/*, */kiiki?u/*, and */?i?ina/*, among others. Most of these words have been transcribed into the International Phonetic Alphabet, though it appears that some words are transcribed using the Roman alphabet. For instance, the second word in the example is likely */iŋko/*, which is a much more likely sequence than what is transcribed.

In this sample, there are 6307 such cognate groups and 210 distinct glosses. The data is somewhat sparse: fewer than 50% of the possible gloss/language pairs are present. Moreover, there is some amount of *homoplasy*—that is, languages with a word from more than one cognate group for a given gloss.

Finally, it is important to note that the ABVD is still a work in progress: they have data from only 50% of extant Austronesian languages.

2.3 Subfamilies of Austronesian

In this paper we focus on two branches of the Austronesian language family, one as a development set and one as a test set. For our development set, we use the Formosan branch. The languages in this group are exclusively found on the Austronesian homeland of Formosa. The family encompasses a substantial portion of the linguistic diversity of Austronesian: Blust (2009) argues that Formosan contains 9 of the 10 first-order splits of the Austronesian family. Formosan's diversity is surprising since it contains a mere 18 languages. Thus, Formosan is a smaller development set that nevertheless is representative of larger families.

For our final test set, we use the Oceanic subfamily, which includes almost 50% of the languages in the Austronesian family, meaning that it represents around 10% of all languages in the world. Oceanic also represents a large fraction of the ge-

ographic diversity of Austronesian, stretching from New Zealand in the south to Hawai'i in the north. Our sample includes 21863 words from 218 languages in the Oceanic family.

3 Models

In this section we describe two models, one based on Hall and Klein (2010)—which we call HK10—and another new model that shares some connection to parsimony methods in computational biology, which we call PARSIM. Both are generative models that describe the evolution of words w_ℓ from a set of languages $\{\ell\}$ in a cognate group g along a fixed phylogeny T.[1] Each cognate group and word is also associated with a gloss or meaning m, which we assume to be fixed.[2] In both models, words evolve according to regular sound laws φ_ℓ, which are specific to each language. Also, both models will make use of a language model λ, which is used for generating words that are not dependent on the word in the parent language. (We leave φ_ℓ and λ as abstract parameters for now. We will describe them in subsequent sections.)

3.1 HK10

The first model we describe is a small modification of the phylogenetic model of Hall and Klein (2010). In HK10, there is an unknown number of cognate groups G where each cognate group g consists of a set of words $\{w_{g,\ell}\}$. In each cognate group, words evolve along a phylogeny, where each word in a language is the result of that word evolving from its parent according to regular sound laws. To model the fact that not all languages have a cognate in each group, each language in the tree has an associated "survival" variable $S_{g,\ell}$, where a word may be lost on that branch (and its descendants) instead of evolving. Once the words are generated, they are then "permuted" so that the cognacy relationships

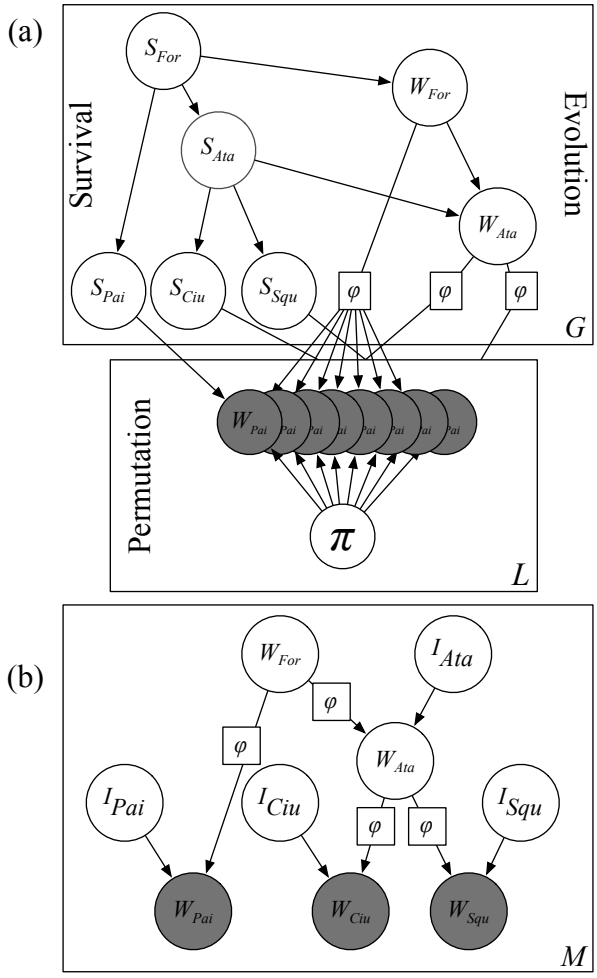

Figure 1: Plate diagrams for (a) HK10 (Hall and Klein, 2010) and (b) PARSIM, our new parsimony model, for a small set of languages. In HK10, words are generated following a phylogenetic tree according to sound laws φ, and then "scrambled" with a permutation π so that the original cognate groups are lost. In PARSIM, all words for each of the M glosses are generated in a single tree, with innovations I starting new cognate groups. The languages depicted are Formosan (For), Paiwan (Pai), Atayalic (Ata), Ciuli Atayalic (Ciu), and Squliq Atayalic (Squ).

[1] Both of these models therefore are insensitive to geographic and historical factors that cannot be easily approximated by this tree. See Nichols (1992) for an excellent discussion of these factors.

[2] One could easily envision allowing the meaning of a word to change as well. Modeling this semantic drift has been considered by Kondrak (2001). In the ABVD, however, any semantic drift has already been elided, since the database has coarsened glosses to the extent that there is no meaningful way to model semantic drift given our data.

are obscured. The task of inference then is to recover the original cognate groups.

The generative process for their model is as follows:

- For each cognate group g, choose a root word $W_{root} \sim p(W|\lambda)$, a language model over words.
- For each language ℓ in a pre-order traversal of the phylogeny:
  1. Choose $S_\ell \sim \text{Bernoulli}(\beta_\ell)$, indicating whether or not the word survives.
  2. If the word survives, choose $W_\ell \sim p(W|\varphi_\ell, W_{\text{par}(\ell)})$.
  3. Otherwise, stop generating words in that language and its descendants.
- For each language, choose a random permutation π of the observed data, and rearrange the cognates according to this permutation.

We reproduce the graphical model for HK10 for a small phylogeny in Figure 1a.

Inference in this model is intractable; to perform inference exactly, one has to reason over all partitions of the data into cognate groups. To address this problem, Hall and Klein (2010) propose an iterative bipartite matching scheme where one language is held out from the others, and then words are assigned to the remaining groups to maximize the probability of the attachment. That is, for some language ℓ and fixed assignments $\pi_{-\ell}$ for the other languages, they seek an assignment π_ℓ that maximizes:

$$\pi^* = \underset{\pi}{\text{argmax}} \sum_g \log p(w_{(\ell,\pi_\ell(g))}|\varphi, \pi, \mathbf{w}_{-\ell})$$

Unfortunately, while this approach was effective with only a few languages (they tested on three), this algorithm cannot scale to the eighteen languages in Formosan, let alone the hundreds of languages in Oceanic. Therefore, we make two simple modifications. First, we restrict the cognate assignments to stay within a gloss. Thus, there are many fewer potential matchings to consider. Second, we use an agglomerative inference procedure, which greedily merges cognate groups that result in the greatest gain in likelihood. That is, for all pairs of cognate groups g_a with words $\mathbf{w}_a$ and g_b with words $\mathbf{w}_b$, we compute the score:

$$\log p(\mathbf{w}_{a\cup b}|\varphi) - \log p(\mathbf{w}_a|\varphi) - \log p(\mathbf{w}_b|\varphi)$$

This score is the difference between the log probability of generating two cognate groups jointly and generating them separately. We then merge the two that generate the highest gain in likelihood. Like the iterative bipartite matching algorithm described above, this algorithm is not exact. However, it is $O(n^2 \log n)$ (where n is the size of the largest gloss, which for Oceanic is 153), while the bipartite matching algorithm is $O(n^3)$ (Kuhn, 1955).

Actually, the original HK10 is doubly intractable. They use weighted automata to represent distributions over strings, but these automata—particularly if they are non-deterministic—make inference in any non-trivial graphical model intractable. We discuss this issue in more detail in Section 6.

3.2 A Parsimony-Inspired Model

We now describe a new model called PARSIM that supports exact inference tractably, though it sacrifices some of the expressive power of HK10. In our model, each language has at most one word for each gloss, and this one word changes from one language to its children according to some edge-specific Markov process. These changes may either be mutations, which merely change the surface form of the word, or innovations, which start a new word in a new cognate group that is unrelated to the previous word. Mutations take the form of a conditional edit operation that models insertions, substitutions, and deletions that correspond to regular (and, with lower probability, irregular) sound changes that are likely to occur between a language and its parent. Innovations, on the other hand, are generated from a language model independent of the parent's word.

Specifically, our generative process takes the following form:

- For each gloss m, choose a root word $W_{root} \sim \lambda$, a language model over words.
- For each language ℓ in a pre-order traversal of the phylogeny:

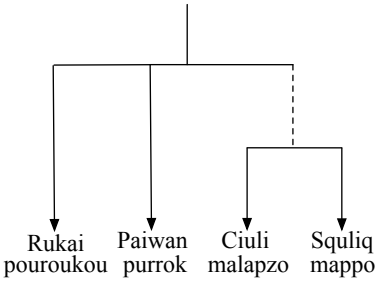

Figure 2: A small example of how PARSIM works. Listed here are the words for "ten" in four languages from the Formosan family, along with the tree that explains them. The dashed line indicates an innovation on the branch.

1. Choose $I_\ell \sim \text{Bernoulli}(\beta_\ell)$, indicating whether or not the word is an innovation or a mutation.
2. If it is a mutation, choose $W_\ell \sim p(W|\varphi_\ell, W_{\text{par}(\ell)})$.
3. Otherwise, choose $W_\ell \sim \lambda$.

We also depict our model as a plate diagram for a small phylogeny in Figure 1b.

Because there is only one tree per gloss, there is no assignment problem to consider, which is the main source of the intractability of HK10. Instead, pieces of the phylogeny are simply "cut" into subtrees whenever an innovation occurs. Thus, message passing can be used to perform inference.

As an example of how our process works, consider Figure 2. The Formosan word for "ten" probably resembled either */purrok/* or */pouroukou/*. There was an innovation in Ciuli and Squliq's ancestor Atayalic that produced a new word for ten. This word then mutated separately into the words */malapzo/* and */mappo/*, respectively.

4 Relation to Parsimony

PARSIM is related to the parsimony principle from computational biology (Cavalli-Sforza and Edwards, 1965; Fitch, 1971), where it is used to search for phylogenies. When using parsimony, a phylogeny is scored according to the derivation that requires the fewest number of changes of state, where a state is typically thought of as a gene or some other trait in a species. These genes are typically called "characters" in the computational biology literature, and two species would have the same value for a character if they share the same property that that state represents.

When inducing phylogenies of languages, a natural choice for characters are glosses from a restricted vocabulary like a Swadesh list, and two words are represented as the same value for a character if they are cognate (Ringe et al., 2002). Other features can be used (Daumé III and Campbell, 2007; Daumé III, 2009), but they are not relevant to our discussion.

Consider the small example in Figure 3a with just four languages. Here, cognacy is encoded using characters. In this example, at least two changes of state are required to explain the data: both C and B must have evolved from A. Therefore, the parsimony score for this tree is two.

Of course, there is no reason why all changes should be equally likely. For instance, it might be extremely likely that B changes into both A and C, but that A never changes into B or C, and so weighted variants of parsimony might be necessary (Sankoff and Cedergren, 1983).

With this in mind, PARSIM can be thought of a weighted variant of parsimony, with two differences. First, the characters do not indicate ahead of time which words are related. Instead, the characters are the words themselves. Second, the transitions between different states (words) are not uniform. Instead, they are weighted by the log probability of one word changing into another, including both mutations and innovations.

Thus, the task of inference in PARSIM is to find the most "parsimonious" explanation for the words we have observed, which is the same as finding the most likely derivation. Because the distances between words (that is, the transition probabilities) are not known ahead of time, they must instead be learned, which we discuss in Section 7.[3]

5 Limitations of the Parsimony Model

Potentially, our parsimony model sacrifices a certain amount of power to make inference tractable. Specifically, it cannot model *homoplasy*, the presence of more than one word in a language for a given

[3]It is worth noting that we are not the first to point out a connection between parsimony and likelihood. Indeed, many authors in the computational biology literature have formally demonstrated a connection (Farris, 1973; Felsenstein, 1973).

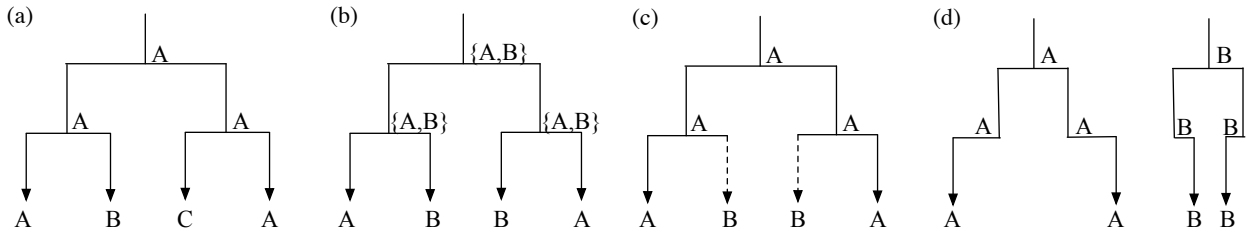

Figure 3: Trees illustrating parsimony and its limitations. In these trees, there are four languages, with words A, B, and C in various configurations. (a) The most parsimonious derivation for this tree has all intermediate states as A. There are thus two changes. (b) An example of homoplasy. Here, given this tree, it seems likely that the ancestral languages contained both A and B. (c) PARSIM cannot recover the example from (b), and so it encodes two innovations (shown as dashed lines). (d) The HK10 model can recover this relationship, but this power makes the model intractable.

gloss. Homoplasy can arise for a variety of reasons in phylogenetic models of cognates, and we describe some in this section.

Consider the example illustrated in Figure 3b, where the two central languages share a cognate, as do the two outer languages. This is the canonical example of homoplasy, and PARSIM cannot correctly recover this grouping. Instead, it can at best only select group A or group B as the value for the parent, and leave the other group fragmented as two innovations, as in Figure 3c. On the other hand, HK10 *can* recover this relationship (Figure 3d), but this power is precisely what makes it intractable.

There are two reasons this kind of homoplasy could arise. The first is that there were indeed two words in the parent language for this gloss, or that there were two words with similar meanings and the two meanings drifted together. Second, the tree could be an inadequate model of the evolution in this case. For instance, there could have been a certain amount of borrowing between two of these languages, or there was not a single coherent parent language, but rather a language continuum that cannot be explained by any tree.

However, homoplasy seems to be relatively uncommon (though not unheard of) in the Oceanic and Formosan families. Where it does appear, our model should simply fail to get one of the cognate groups, instead explaining all of them via innovation. To repair this shortcoming, we can simply run the agglomerative clustering procedure for the model of Hall and Klein (2010), starting from the groups that PARSIM has recovered. Using this procedure, we can hopefully recover many of the under-groupings caused by homoplasy.

6 Inference and Scale

6.1 Inference

In this section we describe the basics of inference in the PARSIM model. We have a nearly tree-structured graphical model (Figure 1); it is not a tree only because of the innovation parameters. Therefore, we apply the common trick of grouping variables to form a tree. Specifically, we group each word variable W_ℓ with its innovation parameter I_ℓ. The distribution of interest is then $p(W_\ell, I_\ell | W_{\mathbf{par}(\ell)}, \phi_\ell, \beta_\ell)$, and the primary operation is summing out messages μ from the children of a language and sending a new message to its parent:

$$\mu_\ell(w_{\mathbf{par}(\ell)}) = \sum_{w_\ell} p(w_\ell|\cdot) \prod_{\ell' \in \mathbf{child}(\ell)} \mu_{\ell'}(w_\ell)$$
$$p(w_\ell|\cdot) = p(w_\ell|I_\ell = 0, w_{\mathbf{par}(\ell)}, \phi_\ell) p(I_\ell = 0|\beta_\ell) + p(w_\ell|I_\ell = 1, \phi_\ell) P(I_\ell = 1|\beta_\ell) \quad (1)$$

The first term involves computing the probability of the word mutating from its parent, and the second involves the probability of the child word from a language model. We describe the parameters and procedures for these operations in 7.1.

6.2 Scale

Even though inference by message-passing in our model is tractable, we needed to make certain concessions to make inference acceptably fast. These choices mainly affect how we represent distributions over strings.

First, we need to model distributions and messages over words on the internal nodes of a phylogeny. The natural choice in this scenario is to use weighted finite automata (Mohri et al., 1996). Automata have been used to successfully model distributions of strings for inferring morphology (Dreyer and Eisner, 2009) as well as cognate detection (Hall and Klein, 2010). Even in models that would be tractable with "ordinary" messages, inference with automata quickly becomes intractable, because the size of the automata grow exponentially with the number of messages passed. Therefore, approximations must be used. Dreyer and Eisner (2009) used a mixture of a k-best list and a unigram language model, while Hall and Klein (2010) used an approximation procedure that projected complex automata to simple, tractable automata using a modified KL divergence.

While either approach could be used here in principle, we found that automata machinery was simply too slow for our application. Instead, we exploit the intuition that we do not need to accurately reconstruct the word for any ancestral language. Moreover, it is inefficient to keep track of probabilities for all strings. Therefore, we only track scores for words that actually exist in a given gloss, which means that internal nodes only have mass on those words. That is, if a gloss has 10 distinct words across all the languages in our dataset, we pass messages that only contain information about those 10 words.

Now, this representation—while more efficient than the automata representations—results in inference that is still quadratic in the number of words in a gloss, since we have distributions of the form $p(w_\ell | w_{\mathrm{par}(\ell)}, \phi_\ell)$. Intuitively, it is unlikely that a word from one distant branch of tree resembles a word in another branch. Therefore, rather than score all of these unlikely words, we use a beam where we only factor in words whose score is at most a factor of e^{-10} less than the maximum score. Our initial experiments found that using a beam provides large savings in time with little impact on prediction quality.

7 Learning

PARSIM has three kinds of parameters that we need to learn: the mutation parameters φ_ℓ, the innovation probabilities β_ℓ, and the global language model λ for generating new words. We learn these parameters via Expectation Maximization (Dempster et al., 1977), iterating between computing expected counts and adjusting parameters to maximize the posterior probability of the parameters. In this section, we describe those parameters.

7.1 Sound Laws

The core piece of our system is learning the sound laws associated with each edge. Since the foundation of historical linguists with the neogrammarians, linguists have argued for the regularity of sound change at the phonemic level (Schleicher, 1861; Bloomfield, 1938). That is to say, if in some language a $/t/$ changes to a $/d/$ in some word, it is almost certain that it will change in every other place that has the same surrounding context.

In practice, of course, sound change is not entirely regular, and complex extralinguistic events can lead to sound changes that are irregular. For example, in some cultures in which Oceanic languages are spoken, the name of the chief is taboo: one cannot speak his name, nor say any word that sounds too much like his name. Speakers of these languages do find ways around this prohibition, often resulting in sound changes that cannot be explained by sound laws alone (Keesing and Fifi'i, 1969).

Nevertheless, we find it useful to model sound change as a largely regular if stochastic process. We employ a sound change model whose expressive power is equivalent to that of Hall and Klein (2010), though with a different parameterization. We model the evolution of a word w_ℓ to its child $w_{\ell'}$ as a sequence of unigram edits that include insertions, deletions, and substitutions. Specifically, we use a standard three-state pair hidden Markov model that is closely related to the classic alignment algorithm of Needleman and Wunsch (1970) (Durbin et al., 2006).

The three states in this HMM correspond to matches/substitutions, insertions, and deletions. The transitions are set up such that insertions and deletions cannot be interleaved. This prevents spurious equivalent alignments, which would cause the model to assign unnecessarily higher probability to transitions with many insertions and deletions.

Actually learning these parameters involves learn-

ing the transition probabilities of this HMM (which model the overall probability of insertion and deletion) as well as the emission probabilities (which model the particular edits). Because there are relatively few words for each language (96 on average in Oceanic), we found it important to tie together the parameters for the various languages, in contrast to Hall and Klein (2010) who did not. In our maximization step, we fit a joint log-linear model for each language, using features that are both specific to a language and shared across languages. Our features included indicators on each substitution, insertion, and deletion operation, along with an indicator for the outcome of each edit operation. This last feature reflects the propensity of a particular phoneme to appear in a given language at all, no matter what its ancestral phoneme was. This parameterization is similar to the one used in the reconstruction system of Bouchard-Côté et al. (2009), except that they used edit operations that conditioned on the context of the surrounding word, which is crucial when trying to accurately reconstruct ancestral word forms. To encourage parameter sharing, we used an ℓ_2 regularization penalty.

7.2 Innovation Parameters

The innovation parameters β_ℓ are parameters for simple Bernoulli distribution that govern the propensity for a language to start a new word. These parameters can be learned separately, though due to data sparsity, we found it better to use a tied parameterization as with the sound laws. Specifically, we fit a log linear model whose features are indicators on the specific language, as well as a global innovation parameter that is shared across all languages. As with the sound laws, we used an ℓ_2 regularization penalty to encourage the use of the global innovation parameter.

7.3 Language Model

Finally, we have a single language model λ that is also shared across all languages. λ is a simple bigram language model over characters in the International Phonetic Alphabet. λ is used when generating new words either via innovation or from the root of the tree.

In principle, we could of course have language models specific to each language, but because there are so few words per language, we found that branch-specific language models caused the model to prefer to innovate at almost every node since the language models could essentially memorize the relatively small vocabularies of these languages.

Formosan				
System	Prec	Recall	F1	Purity
Agg. HK10	77.6	**83.2**	80.0	84.7
PARSIM	**87.8**	71.0	78.5	**94.6**
Combination	85.2	81.3	**83.2**	92.3
Oceanic				
System	Prec	Recall	F1	Purity
PARSIM	**84.4**	62.1	71.5	**91.8**
Combination	76.0	73.8	**74.9**	85.5

Table 1: Results on the Formosan and Oceanic families. PARSIM is the new parsimony model in this paper, Agg. HK10 is our agglomerative variant of Hall and Klein (2010) and Combination uses PARSIM's output to seed the agglomerative matcher. For the agglomerative systems, we report the point with maximal F1 score, but we also show precision/recall curves. (See Figure 4.)

8 Experiments

8.1 Cognate Recovery

We ran both PARSIM and our agglomerative version of HK10 on the Formosan datasets. For PARSIM, we initialized the mutation parameters φ to a model that preferred matches to insertions, substitutions and deletions by a factor of e^3, innovation parameters to 0.5, and the language model to a uniform distribution over characters. For the agglomerative HK10, we initialized its parameters to the values found by our model.[4]

Based on our observations about homoplasy, we also considered a combined system where we ran PARSIM, and then seeded the agglomerative clustering algorithm with the clusters found by PARSIM.

For evaluation, we report a few metrics. First, we report cluster purity, which is a kind of precision measure for clusterings. Specifically, each cluster is assigned to the cognate group that is the most common cognate word in that group, and then purity is computed as the fraction of words that

[4] Attempts to learn parameters directly with the agglomerative clustering algorithm were not effective.

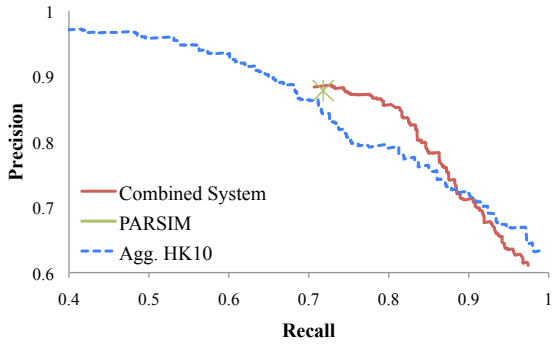

Figure 4: Precision/Recall curves for our systems. The Combined System starts from PARSIM's output, so it has fewer points to plot, and starts from a point with lower precision. As PARSIM outputs only one result, it is starred.

are in a cluster whose gold cognate group matches the cognate group of the cluster. For gold partitions $G = \{G_1, G_2, \ldots, G_g\}$ and found partitions $F = \{F_1, F_2, \ldots, F_f\}$, we have: $\text{purity}(G, F) = \frac{1}{N} \sum_f \max_g |G_g \cap F_f|$. We also report pairwise precision and recall computed over pairs of words.[5] Finally, because agglomerative clustering does not define a natural "stopping point" other than when the likelihood gain decreases to 0—which did not perform well in our initial tests—we will report both a precision/recall curve, as well the maximum pairwise F1 obtained by the agglomerative HK10 and the combined system.

The results are in Table 1. On Formosan, PARSIM has much higher precision and purity than our agglomerative version of HK10 at its highest point, though its recall and F1 suffer somewhat. Of course, the comparison is not quite fair, since we have selected the best possible point for HK10.

However, our combination of the two systems does even better. By feeding our high-precision results into the agglomerative system and sacrificing just a little precision, our combined system achieves much higher F1 scores than either of the systems alone.

Next, we also examined precision and recall curves for the two agglomerative systems on Formosan, which we have plotted in Figure 4, along with the one point output by PARSIM.

We then ran PARSIM and the combined system on the much larger Oceanic dataset. Performance on all metrics decreased somewhat, but this is to be expected since there is so much more data. As with Formosan, PARSIM has higher precision than the combined system, but it has much lower recall.

8.2 Reconstruction

We also wanted to see how well our cognates could be used to actually reconstruct the ancestral forms of words. To do so, we ran a version of Bouchard-Côté et al. (2009)'s reconstruction system using both the cognate groups PARSIM found in the Oceanic language family and the gold cognate groups provided by the ABVD. We then evaluated the average Levenshtein distance of the reconstruction for each word to the reconstruction of that word's Proto-Oceanic ancestor provided by linguists. Our evaluation differs from Bouchard-Côté et al. (2009) in that they averaged over cognate groups, which does not make sense for our task because there are different cognate groups. Instead, we average over per-modern-word reconstruction error.

Using this metric, reconstructions using our system's cognates are an average of 2.47 edit operations from the gold reconstruction, while with gold cognates the error is 2.19 on average. This represents an error increase of 12.8%. To see if there was some pattern to these errors, we also plotted the fraction of words with each Levenshtein distance for these reconstructions in Figure 5. While the plots are similar, the automatic cognates exhibit a longer tail. Thus, even with automatic cognates, the reconstruction system can reconstruct words faithfully in many cases, but in a few instances our system fails.

9 Analysis

We now consider some of the errors made by our system. Broadly, there are two kinds of mistakes in a model like ours: those affecting precision and those affecting recall.

9.1 Precision

Many of our precision errors seem to be due to our somewhat limited model of sound change. For instance, the language Pazeh has two words for

[5]The main difference between precision and purity is that pairwise precision is inherently quadratic, meaning that it penalizes mistakes in large groups much more heavily than mistakes in small groups.

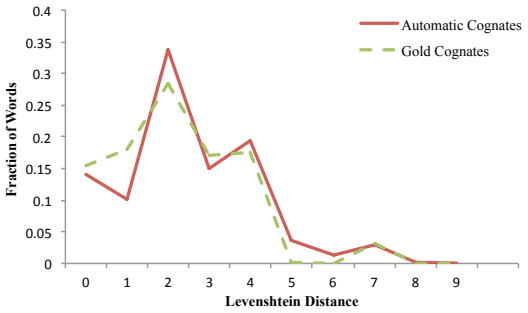

Figure 5: Percentage of words with varying levels of Levenshtein distance from the gold reconstruction. Gold Cognates were hand-annotated by linguists, while Automatic Cognates were found by our system.

"to sleep:" */mudamai/* and */midəm/*. Somewhat surprisingly the former word is cognate with Paiwan */qmereŋ/* and Saisiat */maʔrəm/* while the latter is not. Our system, however, makes the mistake of grouping */midəm/* with the Paiwan and Saisiat words. Our system has inferred that the insertions of */u/* and */ai/* (which are required to bring */mudamai/* into alignment with the Saisiat and Paiwan words) are less likely than substituting a few vowels and the consonant */r/* for */d/* (which are required to align */midəm/*). Perhaps a more sophisticated model of sound change could correctly learn this relationship.

However, a preliminary inspection of the data seems to indicate that not all of our precision errors are actually errors, but rather places where the data is insufficiently annotated (and indeed, the ABVD is still a work in progress). For instance, consider the words for "meat/flesh" in the Formosan languages: Squliq */hiʔ/*, Bunun */titiʔ/*, Paiwan */seti/*, Kavalan */ʔisiʔ/*, CentralAmi */titi/*, Our system groups all of these words except for Squliq */hiʔ/*. However, despite these words' similarity, there are actually three cognate groups here. One includes Squliq */hiʔ/* and Kavalan */ʔisiʔ/*, another includes just Paiwan */seti/*, and the third includes Bunun */titiʔ/* and CentralAmi */titi/*. Crucially, these cognate groups do not follow the phylogeny closely. Thus, either there was a significant amount of borrowing between these languages, or there was a striking amount of homoplasy in Proto-Formosan, or these words are in fact mostly cognate. While a more thorough, linguistically-informed analysis is needed to ensure that these are actually cognates, we believe that our system, in conjunction with a trained Austronesian specialist, could potentially find many more cognate groups, speeding up the process of completing the ABVD.

9.2 Recall

Our system can also fail to group words that should be grouped. One recurring problem seems to be reduplication, which is a fairly common phenomenon in Austronesian languages. For instance, there is a cognate group for "to eat" that includes Bunun */maun/*, Thao */kman/*, Favorlang */man/*, and Sediq */manakamakan/*, among others. Our system correctly finds this group, with the exception of */manakamakan/*, which is clearly the result of reduplication. Reduplication cannot be modeled using mere sound laws, and so a more complex transition model is needed to correctly identify these kinds of changes.

10 Conclusion

We have presented a new system for automatically finding cognates across many languages. Our system is comprised of two parts. The first, PARSIM, is a new high-precision generative model with tractable inference. The second, HK10, is a modification of Hall and Klein (2010) that makes their approximate inference more efficient. We discuss certain trade-offs needed to make both models scale, and demonstrated its performance on the Formosan and Oceanic language families.

Acknowledgments

The authors would like to thank to Alexandre Bouchard for providing data. This work was supported by the NSF under grant 1018733, and by an NSF fellowship to the first author.

References

Shane Bergsma and Greg Kondrak. 2007. Multilingual cognate identification using integer linear programming. In *RANLP Workshop on Acquisition and Management of Multilingual Lexicons*, Borovets, Bulgaria, September.

Leonard Bloomfield. 1938. *Language*. Holt, New York.

R. A. Blust. 2009. *The Austronesian languages*. Australian National University.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *EMNLP*.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *NAACL*, pages 65–73.

L. L. Cavalli-Sforza and A. W. F. Edwards. 1965. Analysis of human evolution. In S. J. Geerts Genetics Today, editor, *Proceedings of XIth International Congress of Genetics, 1963, Vol*, page 923–933. 3, 3.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *NAACL*.

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.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *EMNLP*, Singapore, August.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. 2006. *Biological sequence analysis*. eleventh edition.

James S. Farris. 1973. On Comparing the Shapes of Taxonomic Trees. *Systematic Zoology*, 22(1):50–54, March.

J. Felsenstein. 1973. Maximum likelihood and minimum steps methods for estimating evolutionnary trees from data on discrete characters. *Systematic Zoology*, 23:240–249.

W. M. Fitch. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology*, 20:406–416.

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

David Hall and Dan Klein. 2010. Finding cognates using phylogenies. In *Association for Computational Linguistics (ACL)*.

Robert M. Keesing and Jonathan Fifi'i. 1969. Kwaio word tabooing in its cultural context. *Journal of the Polynesian Society*, 78(2):154–177.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *NAACL*.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

John B. Lowe and Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20(3):381–417.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *NAACL*.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 1996. Weighted automata in text and speech processing. In *ECAI-96 Workshop*. John Wiley and Sons.

Andrea Mulloni. 2007. Automatic prediction of cognate orthography using support vector machines. In *ACL*, pages 25–30.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

John Nerbonne. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

J. Nichols. 1992. *Linguistic diversity in space and time*. University of Chicago Press.

Michael P. Oakes. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Quantitative Linguistics*, 7(3):233–243.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-european and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

D. Sankoff and R. J. Cedergren, 1983. *Simultaneuous comparison of three or more sequences related by a tree*, page 253–263. Addison-Wesley, Reading, MA.

August Schleicher. 1861. *A Compendium of the Comparative Grammar of the Indo-European, Sanskrit, Greek and Latin Languages.*