Continuous Space Representations of Linguistic Typology and their Application to Phylogenetic Inference

Yugo Murawaki
Graduate School of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
murawaki@ait.kyushu-u.ac.jp

Abstract

For phylogenetic inference, linguistic typology is a promising alternative to lexical evidence because it allows us to compare an arbitrary pair of languages. A challenging problem with typology-based phylogenetic inference is that the changes of typological features over time are less intuitive than those of lexical features. In this paper, we work on reconstructing typologically natural ancestors To do this, we leverage dependencies among typological features. We first represent each language by continuous latent components that capture feature dependencies. We then combine them with a typology evaluator that distinguishes typologically natural languages from other possible combinations of features. We perform phylogenetic inference in the continuous space and use the evaluator to ensure the typological naturalness of inferred ancestors. We show that the proposed method reconstructs known language families more accurately than baseline methods. Lastly, assuming the monogenesis hypothesis, we attempt to reconstruct a common ancestor of the world's languages.

1 Introduction

Linguistic typology is a cross-linguistic study that classifies the world's languages according to structural properties such as complexity of syllable structure and object-verb ordering. The availability of a large typology database (Haspelmath et al., 2005) makes it possible to take computational approaches to this area of study (Daumé III and Campbell, 2007; Georgi et al., 2010; Rama and Kolachina, 2012). In this paper, we consider its application to phylogenetic inference. We aim at reconstructing evolutionary trees that illustrate how modern languages have descended from common ancestors.

Typological features have two advantages over other linguistic traits. First, they allow us to compare an arbitrary pair of languages. By contrast, historical linguistics has worked on regular sound changes (see (Bouchard-Côté et al., 2013) for computational models). Glottochronology and computational phylogenetics make use of the presence and absence of lexical items (Swadesh, 1952; Gray and Atkinson, 2003). All these approaches require that certain sets of cognates, or words with common etymological origins, are shared by the languages in question. For this reason, it is hardly possible to use lexical evidence to search for external relations involving language isolates and tiny language families such as Ainu, Basque, and Japanese. For these languages, typology can be seen as the last hope.

The second advantage is that typological features are potentially capable of tracing evolutionary history on the order of 10,000 years because they change far more slowly than lexical traits. A glottochronological study indicates that even if Japanese is genetically related to Korean, they diverged from a common ancestor no earlier than 6,700 years ago (Hattori, 1999). Even the basic vocabulary vanishes so rapidly that after some 6,000 years, the retention rate becomes comparable to chance similarity. By contrast, the word order of Japanese, for example is astonishingly stable. It remains intact from the earliest attested data. Thus we argue that if we manage to develop a statistical model of typological

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 324–334,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

	Munda	**Mon-Khmer**
grammar	synthetic	analytic
word order	head-last, OV, postpositional	head-first, VO, prepositional
affixation	pre/infixing, suffixing	pre/infixing or isolating
fusion	agglutinative	fusional
consonants	stable/assimilative	shifting/dissimilative
vowels	harmonizing/stable	reducing/diphthongizing

Table 1: Typological comparison of the Munda and Mon-Khmer branches of the Austroasiatic languages. An abridged version of Table 1 of (Donegan and Stampe, 2004).

changes with predictive power, we can understand a much deeper past.

A challenging problem with typology-based inference is that the changes of typological features over time are less intuitive than those of lexical features. Regular sound changes have been well known since the time of the Neogrammarians. The binary representations of lexical items commonly used in computational phylogenetics correspond to their their presence and absence. The alternations of each feature value can be straightforwardly interpreted as the birth and death (Le Quesne, 1974) of a lexical item. By contrast, it is difficult to understand how a language switches from SOV to SVO.

Practically speaking, since each language is represented by a vector of categorical features, we can easily perform distance-based hierarchical clustering. Still, the extent to which the resultant tree reflects evolutionary history is unclear. Teh et al. (2008) proposed a generative model for hierarchical clustering, which straightforwardly explains evolutionary history. However, features used in their experiments were binarized in a one-versus-rest manner (i.e., expanding a feature with K possible values into K binary features) (Daumé III and Campbell, 2007) although the model itself had an ability to handle categorical values. With the independence assumption of binary features, the model was likely to reconstruct ancestors with logically impossible states.

Typological studies have shown that dependencies among typological features are not limited to the categorical constraints. For example, object-verb ordering is said to imply adjective-noun ordering (Greenberg, 1963). A natural question arises as to what would happen to adjective-noun ordering if object-verb ordering were altered. While dependencies among feature *pairs* were discussed in previous studies (Greenberg, 1978; Dunn et al., 2011), dependencies among more than two features are yet to be exploited.

To gain a better insight into typological changes, we take Austroasiatic languages as an example. Table 1 compares some typological features of the Munda and Mon-Khmer branches. Although their genetic relationship was firmly established, they are almost opposite in structure. Their common ancestor is considered to have been Mon-Khmer-like. This indicates that the holistic changes have happened in the Munda branch (Donegan and Stampe, 2004). To generalize from this example, we suggest the following hypotheses:

1. The holistic polarization can be explained by latent components that control dependencies among observable features.
2. Typological changes can occur in a way such that typologically unnatural intermediate states are avoided.

To incorporate these hypotheses, we propose continuous space representations of linguistic typology. Specifically, we use an autoencoder (see (Bengio, 2009) for a review) to map each language into the latent space. In analogy with principal component analysis (PCA), each element of the encoded vector is referred to as a component. We combine the autoencoder with a typology evaluator that distinguishes typologically natural languages from other possible combinations of features.

Armed with the typology evaluator, we perform phylogenetic inference in the continuous space. The evaluator ensures that inferred ancestors are also typologically natural. The inference procedure is guided by known language families so that each component's stability with respect to evolutionary history can be learned. To evaluate the proposed method, we hide some trees to see how well they are reconstructed.

Lastly, we build a binary tree on top of known language families. This experiment is based on a controversial assumption that the world's languages descend from one common ancestor. Our goal here is not to address the validity of the monogenesis hypothesis. Rather, we address the questions of how the common ancestor looked like if it existed and how modern languages have evolved from it.

2 Related Work

In linguistic typology, much attention has been given to non-tree-like evolution (Trubetzkoy, 1928). Daumé III (2009) incorporated linguistic areas into a phylogenetic model and reported that the extended model outperformed a simple tree model. This result motivates us to use known language families for supervision rather than to perform phylogenetic inference in purely unsupervised settings.

Dunn et al. (2011) applied a state-process model to reference phylogenetic trees to test if a pair of features is independent. The model they adopted can hardly be extended to handle multiple features. They separately applied the model to each language family and claimed that most dependencies were lineage-specific rather than universal tendencies. However, each known language family is so shallow in time depth that few feature changes can be observed in it (Croft et al., 2011). We mitigate data sparsity by letting our model share parameters among language families all over the world.

3 Data and Preprocessing

3.1 Typology Database and Phylogenetic Trees

The typology database we used is the *World Atlas of Language Structures* (WALS) (Haspelmath et al., 2005). As of 2014, it contains 2,679 languages and 192 typological features. It covers less than 15% of the possible language/feature pairs, however.

WALS provides phylogenetic trees but they only have two layers above individual languages: family and genus. Language families include Indo-European, Austronesian and Niger-Congo, and genera within Indo-European include Germanic, Indic and Slavic. For more detailed trees, we used hierarchical classifications provided by Ethnologue (Lewis et al., 2014). The mapping between WALS and Ethnologue was done using ISO 639-3 language codes. We manually corrected some obsolete language codes used by WALS and dropped languages without language codes. We also excluded languages labeled by Ethnologue as Deaf sign language, Mixed language, Creole or Unclassified. For both WALS and Ethnologue trees, we removed intermediate nodes that had only one child. Language isolates were treated as family trees of their own. We obtained 193 family trees for WALS and 189 for Ethnologue.

We made no further modifications to the trees although we were aware that some language families and their subgroups were highly controversial. In the future work, the Altaic language family, for example, should be disassembled into Turkic, Mongolic and Tungusic to test if the Altaic hypothesis is valid (Vovin, 2005).

Next, we removed features with low coverage. Some features such as "Inclusive/Exclusive Forms in Pama-Nyungan" (39B) and "Irregular Negatives in Sign Languages" (139A) were not supposed to cover the world. We selected 98 features that covered at least 10% of languages.[1]

We used the original, categorical feature values. The mergers of some fine-grained feature values seem desirable (Daumé III and Campbell, 2007; Greenhill et al., 2010; Dunn et al., 2011). Some features like "Consonant Inventories" might be better represented as real-valued features. We leave them for future work.

In the end, we created two sets of data. The first set PARTIAL was used to train the typology evaluator. We selected 887 languages that covered at least 30% of features. The second set FULL was for phylogenetic inference. We chose language families in each of which at least 30% of features were covered by one or more languages in the family. The numbers of language families (including language isolates) were reduced to 103 for WALS and 110 for Ethnologue.

3.2 Missing Data Imputation

We imputed missing data using the R package *missMDA* (Josse et al., 2012). It handled missing values using multiple correspondence analysis (MCA). Specifically, we used the imputeMCA function to

[1] Additional cleanup is needed. For example, the high-coverage feature "The Position of Negative Morphemes in SOV Languages" (144L) is not defined for non-SOV languages. A natural solution is to add another feature value (*Undefined*).

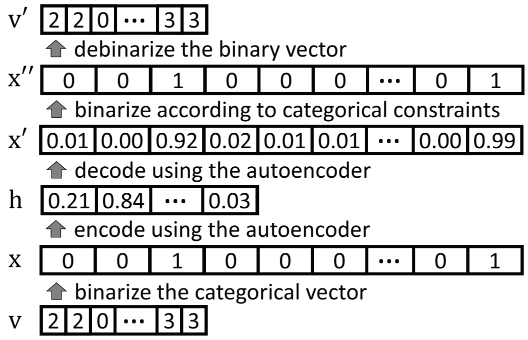

v′ 2 2 0 ··· 3 3
debinarize the binary vector
x″ 0 0 1 0 0 0 ··· 0 1
binarize according to categorical constraints
x′ 0.01 0.00 0.92 0.02 0.01 0.01 ··· 0.00 0.99
decode using the autoencoder
h 0.21 0.84 ··· 0.03
encode using the autoencoder
x 0 0 1 0 0 0 ··· 0 1
binarize the categorical vector
v 2 2 0 ··· 3 3

Figure 1: Representations of a language.

predict missing feature values. The substituted data are used (1) to train the typology evaluator and (2) to initialize phylogenetic inference.

To evaluate the performance of missing data imputation, we hid some known features to see how well they were predicted. A 10-fold cross-validation test using the PARTIAL dataset showed that 64.6% of feature values were predicted correctly. It considerably outperformed (1) the random baseline of 22.4% and (2) the most-frequent-value baseline of 28.1%. Thus our assumption of dependencies among features was confirmed.

4 Typology Evaluator

We use a combination of an autoencoder to transform typological features into continuous latent components, and an energy-based model to evaluate how a given feature vector is typologically natural.

We begin with the autoencoder. Figure 1 shows various representations of a language. The original feature representation v is a vector of categorical features. v is binarized into $\mathrm{x} \in \{0,1\}^{d_0}$ in a one-versus-rest manner. x is mapped by an encoder to a latent representation $\mathrm{h} \in [0,1]^{d_1}$, in which d_1 is the dimension of the latent space:

$$\mathrm{h} = s(\mathrm{W_e}\mathrm{x} + \mathrm{b_e}),$$

where s is the sigmoid function, and matrix $\mathrm{W_e}$ and vector $\mathrm{b_e}$ are weight parameters to be estimated. A decoder then maps h back to x' through a similar transformation:

$$\mathrm{x}' = s(\mathrm{W_d}\mathrm{h} + \mathrm{b_d}).$$

We use tied weights: $\mathrm{W_d} = \mathrm{W_e}^T$. Note that x' is a real vector. To recover a categorical vector, we need to first binarize x' according to categorical constraints and then to debinarize the resultant vector.

The training objective of the autoencoder alone is to minimize cross-entropy of reconstruction:

$$L_{\mathrm{AE}}(\mathrm{x}, \mathrm{x}') = -\sum_{k=1}^{d} x_k \log x'_k + (1-x_k)\log(1-x'_k),$$

where x_k is the k-th element of x.

Next, we plug an energy-based model into the autoencoder. It gives a probability to x.

$$\begin{aligned} p(\mathrm{x}) &= \frac{\exp(\mathrm{W_s}^T \mathrm{g})}{\sum_{\mathrm{x}'} \exp(\mathrm{W_s}^T \mathrm{g}')}, \\ \mathrm{g} &= s(\mathrm{W_l}\mathrm{h} + \mathrm{b_l}), \end{aligned}$$

where vector $\mathrm{W_s}$, matrix $\mathrm{W_l}$ and bias term $\mathrm{b_l}$ are the weights to be estimated. h is mapped to $\mathrm{g} \in [0,1]^{d_2}$ before evaluation. This transformation is motivated by our speculation that typologically natural languages may not be linearly separable from unnatural ones in the latent space since biplots of principal components of PCA often show sinusoidal waves (Novembre and Stephens, 2008). The denominator sums over all possible states of x', including those which violate categorical constraints. By maximizing the average log probability of training data, we can distinguish typologically natural languages from other possible combinations of features.

Given a set of N languages with missing data imputed,[2] our training objective is to maximize the following:

$$\sum_{i=1}^{N} (-L_{\mathrm{AE}}(\mathrm{x}_i, \mathrm{x}'_i) + C \log p(\mathrm{x}_i))),$$

where C is some constant. Weights are optimized by the gradient-based AdaGrad algorithm (Duchi et al., 2011) with a mini-batch. A problem with this optimization is that the derivative of the second term contains an expectation that involves a summation over all possible states of x', which is computationally intractable. Inspired by contrastive divergence (Hinton, 2002), we do not compute the expectation exactly but approximate it by few negative samples collected from Gibbs samplers.

4.1 Mixing Languages: An Experiment

To analyze the continuous space representations, we generated mixtures of two languages, which were

[2] We tried a joint inference of weight optimization and missing data imputation but dropped it for its instability. A cross-validation test revealed that the joint inference caused a big accuracy drop in missing data imputation.

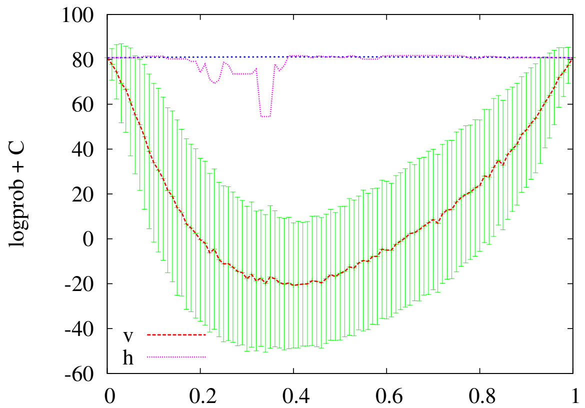


Figure 2: Mixtures of Mundari (a Munda language) and Khmer (a Mon-Khmer language). The transitions from Mundari (leftmost) to Khmer (rightmost). The vertical axis denotes typological naturalness $\log p(\mathrm{x}) + C$.

potential candidates for their common ancestor. The pair of languages A and B was mixed in two ways. First, we replaced elements of A's categorical vector v_A with v_B, with the specified probability. We repeated this procedure 1,000 times to obtain a mean and a standard deviation. Second, we applied linear interpolation of two vectors h_A and h_B and mapped the resultant vector to v'. In this experiment, $d_0 = 539$ and we set $d_1 = 100$ and $d_2 = 10$.

Figure 2 shows the case of the Austroasiatic languages. In the original, categorical representations, the mixtures of two languages form a deep valley (i.e., typologically unnatural intermediate states). By contrast, the continuous space representations allow a language to change into another without harming typological naturalness. This indicates that in the continuous space, we can easily reconstruct typologically natural ancestors. The major feature changes include "postpositional" to "prepositional" (0.46–0.47), "strongly suffixing" to "little affixation" (0.53–0.54) and "SOV" to "SVO" (0.60–0.61).

5 Phylogenetic Inference

5.1 Tree Model

We use continuous space representations and the typology evaluator for phylogenetic inference. Our strategy is to find a tree in which (1) nodes are typologically natural and (2) edges are shorter by the principle of Occam's razor. The first point is realized by applying the typology evaluator. To implement the second point, we define a probability distribution over a parent-to-child move in the continuous space.

We assume that latent components are independent. For the k-th component, the node's value h_k is drawn from a Normal distribution with mean h_k^{P} (its parent's value) and precision λ_k (inverse variance). The further the node moves, the smaller probability it receives. Precision controls each component's stability with respect to evolutionary history.

We set a gamma prior over λ_k, with hyperparameters α and β.[3] Taking advantage of the conjugacy property, we marginalize out λ_k. Suppose that we have drawn n samples and let m_i be the difference between the i-th node and its parent, $h_k - h_k^{\mathrm{P}}$. Then the posterior hyperparameters are $\alpha_n = \alpha + n/2$ and $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n} m_i^2$. The posterior predictive distribution is Student's t-distribution (Murphy, 2007):

$$p_k(h_k | h_k^{\mathrm{P}}, M_{\mathrm{hist}}, \alpha, \beta) = t_{2\alpha_n}(h_k | h_k^{\mathrm{P}}, \sigma^2 = \beta_n/\alpha_n),$$

where M_{hist} is a collection of α, β and a history of previously observed differences. The probability of a parent-to-child move is a product of the probabilities of its component moves:

$$p_{\textsc{Move}}(\mathrm{h}|\mathrm{h}^{\mathrm{P}}, M_{\mathrm{hist}}) = \prod_{k=1}^{d} p_k(h_k | h_k^{\mathrm{P}}, M_{\mathrm{hist}}).$$

The root node is drawn from a uniform distribution.

To sum up, the probability of a phylogenetic tree $\boldsymbol{\tau}$ is given by $p_{\textsc{Eval}}(\text{tree}) \times p_{\textsc{Cont}}(\text{tree})$, where

$$\begin{aligned} p_{\textsc{Eval}}(\text{tree}) &= \text{Uniform}(\text{tree}) \prod_{\mathrm{x} \in \text{nodes}(\boldsymbol{\tau})} p(\mathrm{x}), \\ p_{\textsc{Cont}}(\text{tree}) &= \text{Uniform}(\text{root}) \\ &\times \prod_{(\mathrm{h}, \mathrm{h}^{\mathrm{P}}) \in \text{edges}(\boldsymbol{\tau})} p_{\textsc{Move}}(\mathrm{h}|\mathrm{h}^{\mathrm{P}}, M_{\mathrm{hist}}). \end{aligned}$$

nodes($\boldsymbol{\tau}$) is the set of nodes in $\boldsymbol{\tau}$, and edges($\boldsymbol{\tau}$) is the set of edges in $\boldsymbol{\tau}$, We abuse notation as M_{hist} is updated each time a node is observed.

5.2 Inference

Given observed data, we aim at reconstructing the best phylogenetic tree. The data observed are (1) leaves (with some missing feature values) and (2) some tree topologies. We need to infer (1) the missing feature values of leaves, (2) the latent components of internal nodes including the root and (3) the remaining portion of tree topologies. Since leaves

[3] In the experiments, we set $\alpha = \beta = 0.1$.

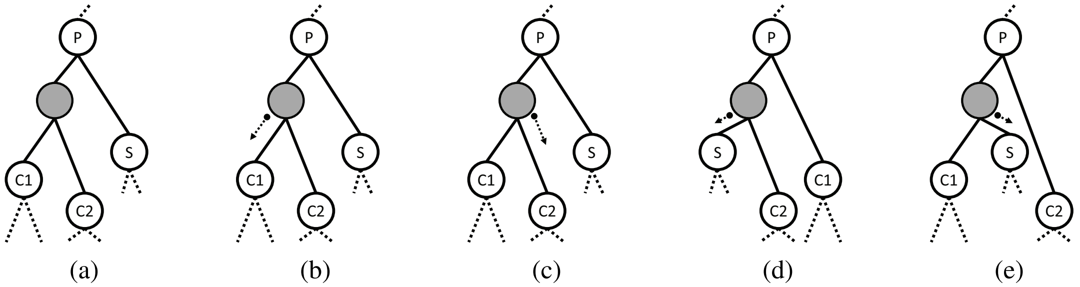


Figure 3: SWAP operator. The gray circle is the target node. Its parent P, sibling S and two children C1 and C2 are shown. (a) The current state. (b–e) The proposed states. (b–c) The topology remains the same but the target is moved toward C1 and C2, respectively. (d) C1 is swapped for S. (e) C2 is swapped for S.

are tied to observed categorical vectors, our inference procedures also work on them. We map categorical vectors into the latent space every time we attempt to change a feature value. By contrast, we adopt latent vectors as the primary representations of internal nodes.

Take the Indo-European language family for example. Its tree topology is given but the states of its internal nodes such as Indo-European, Germanic and Indic need to be inferred. Dutch has some missing feature values. Although they have been imputed with multiple correspondence analysis, its close relatives such as Danish and German might be helpful for better estimation.

We need to infer portions of tree topologies even though a set of trees (language families) is given. To evaluate the performance of phylogenetic inference, we hide some trees to see how well they are reconstructed. To reconstruct a common ancestor of the world's languages, we build a binary tree on top of the set of trees. Note that while we only infer binary trees, a node may have more than two children in the fixed portions of tree topologies.

We use Gibbs sampling for inference. We define four operators, CAT, COMP, SWAP and MOVE. The first tree operators correspond to missing feature values, latent components and tree topologies, respectively.

CAT – For the target categorical feature of a leaf node, we sample from K possible values. Let x' be a binary feature representation with the target feature value altered, let h^{P} be the state of the node's parent, and let $\mathrm{h}' = s(\mathrm{W}_{\mathrm{e}}\mathrm{x}' + \mathrm{b}_{\mathrm{e}})$. The probability of choosing x' is proportional to $p(\mathrm{x}')\, p_{\textsc{Move}}(\mathrm{h}'|\mathrm{h}^{\mathrm{P}}, M_{\mathrm{hist}})$, where h is removed from the history. The second term is omitted if the target node has no parent.[4]

COMP – For the target k-th component of an internal node, we choose its new value using the Metropolis algorithm. It stochastically proposes a new state and accepts it with some probability. If the proposal is rejected, the current state is reused as the next state. The proposal distribution $Q(h'_k|h_k)$ is a Gaussian distribution centered at h_k. The acceptance probability is $a(h_k, h'_k) = \min(1, P(h'_k)/P(h_k))$, where $P(h'_k)$ is defined as

$$P(h'_k) = p(\mathrm{x}')\, p_{\textsc{Move}}(\mathrm{h}'|\mathrm{h}^{\mathrm{P}}, M_{\mathrm{hist}}) \prod_{\mathrm{h}^{\mathrm{C}} \in \mathrm{children}(\mathrm{h}')} p_{\textsc{Move}}(\mathrm{h}^{\mathrm{C}}|\mathrm{h}', M_{\mathrm{hist}})$$

where $\mathrm{children}(\mathrm{h}')$ is the set of the target node's children.

SWAP – For the target internal node (which cannot be the root), we use the Metropolis-Hastings algorithm to locally rearrange its neighborhood in a way similar to Li et al. (2000). We first propose a new state as illustrated in Figure 3. The target node has a parent P, a sibling S and two children C1 and C2. From among S, C1 and C2, we choose two nodes. If C1 and C2 are chosen, the topology remains the same; otherwise S is swapped for one of the node's children. It is shown that one topology can be transformed into any other topology in a finite number of steps (Li et al., 2000).

To improve mobility, we also move the target node toward C1, C2 or S, depending on the proposed topology. Here the selected node is denoted by $*$. We first draw r' from a log-normal distribution whose underlying Gaussian distribution has

[4] It is easy to extend the operator to handle internal nodes supplied with some categorical features.

mean -1 and variance 1. The target's proposed state is $\mathrm{h}' = (1 - r')\mathrm{h} + r'\mathrm{h}^*$. r' can be greater than 1, and in that case, the proposed state h' is more distant from h^* than the current state h. This ensures that the transition is reversible because $r = 1/r'$. The acceptance probability can be calculated in a similar manner to that described for Comp.

Move – Propose to move the target internal node, without swapping its neighbors.

For initialization, missing feature values are imputed by *missMDA*. The initial tree is constructed by distance-based agglomerative clustering. The state of an internal node is set to the average of those of its children.

6 Experiments

6.1 Reconstruction of Known Family Trees

6.1.1 Data and Method

We first conducted a quantitative evaluation of phylogenetic inference, using known family trees. We ran 5-fold cross-validations. For each of WALS and Ethnologue, we subdivided a set of language families into 5 subsets with roughly the same number of leaves. Because of some huge language families, the number of language families per subset was uneven. We disassembled family trees in the target subset and to let the model reconstruct a binary tree for each language family. Unlike ordinary held-out evaluation, this experiment used all data for inference at once.

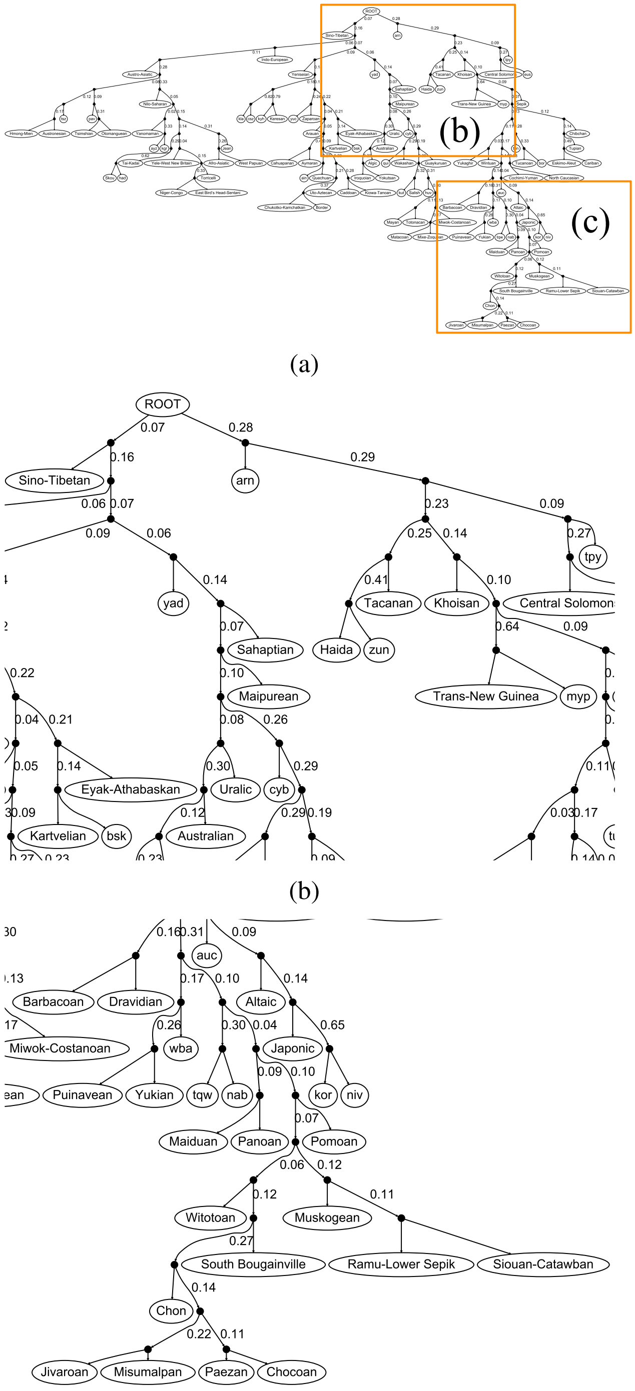
6.1.2 Model Settings

We used the parameter settings described in Section 4.1. For phylogenetic inference, we ran 9,000 burn-in iterations after which we collected 100 samples at an interval of 10 iterations.

For comparison, we performed average-link agglomerative clustering (Alc). It has two variants, Alc-Cat and Alc-Cont. Alc-Cat worked on categorical features and used the ratio of disagreement as a distance metric. Alc-Cont performed clustering in the continuous space, using cosine distance. In other words, we can examine the effects of the typology evaluator and precision parameters. For these models, missing feature values are imputed by *missMDA*.

6.1.3 Evaluation Measures

We present purity (Heller and Ghahramani, 2005), subtree (Teh et al., 2008) and outlier fraction scores (Krishnamurthy et al., 2012). All scores are between 0 and 1 and higher scores are better. We calculated these scores for each language family and



(a)

(b)

(c)

Figure 4: Maximum clade credibility tree of the world. (a) The whole tree. Three-letter labels are ISO 639-3 codes. Nodes below language families are omitted. (b–c) Portions of the tree are enlarged.

	WALS						Ethnologue	
	purity		subtree		outlier		outlier	
Alc-Cat	.500	.557	.608	.626	.343	.330	**.358**	**.398**
Alc-Cont	.503	.557	**.630**	.630	.343	.330	.353	.395
Proposed	**.522**	**.572**	.603	**.651**	**.351**	**.346**	.356	.394

Table 2: Results of the reconstruction of known family trees. Macro-averages are followed by micro-averages.

report macro- and micro-averages. Only non-trivial family trees (trees with more than two children) were considered.

Purity and subtree scores compare inferred trees with gold-standard class labels. In WALS, genera were treated as class labels because they were the only intermediate layer between families and leaves. By contrast, Ethnologue provided more complex trees and we were unable to assign one class label to each language. For this reason, only outlier fraction scores are reported for Ethnologue.

6.1.4 Results

Table 2 shows the scores for reconstructed family trees. The proposed method outperformed the baselines in 5 out of 8 metrics. Three methods performed almost equally for Ethnologue. We suspect that typological features reflect long term trends in comparison to Ethnologue's fine-grained classification. For WALS, the proposed method was beaten by average-link agglomerative clustering only in the macro-average of subtree scores. One possible explanation is randomness of the proposed method. Apparently, random sampling distributed errors more evenly than deterministic clustering. It was penalized more often by subtree scores because they required that all leaves of an internal node belonged to the same class.

6.2 Reconstruction of a Common Ancestor of the World's Languages

We reconstructed a single tree that covers the world. To do this, we build a binary tree on top of known language families, a product of historical linguistics. It is generally said that historical linguistics cannot go far beyond 6,000–7,000 years (Nichols, 2011). Here we attempt to break the brick wall.

It is no surprise that this experiment is full of problems and difficulties. No quantitative evaluation is possible. Underlying assumptions are questionable. No one knows for sure if there was such a thing as one common ancestor of all modern languages. Moreover, *language capacity* of humans, in addition to languages themselves, is likely to have evolved over time (Nichols, 2011). This casts doubt on the applicability of the typology evaluator, which is trained on modern languages, to languages of far distant past. Nevertheless, it is fascinating to make inference on the world's ancestral languages.

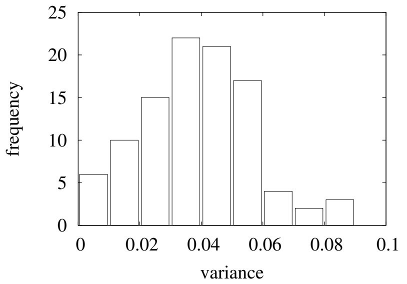


Figure 5: Histogram of posterior variances $\sigma^2 = \beta_n/\alpha_n$ of the 4,000th sample.

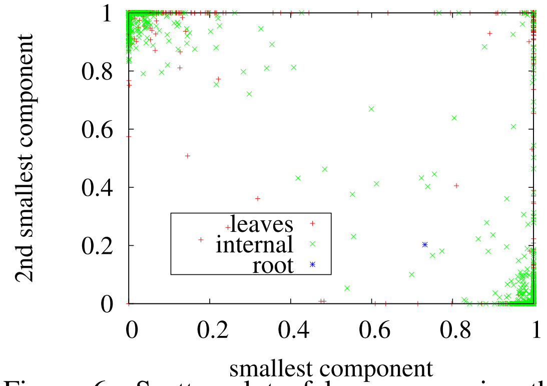


Figure 6: Scatter plot of languages using the components with the two smallest variances.

We used Ethnologue as the known tree topologies. For Gibbs sampling, we ran 3,000 burn-in iterations after which we collected 100 samples at an interval of 10 iterations.

Figure 4 shows a reconstructed tree. To summarize multiple sample trees, we constructed a maximum clade credibility tree. For each clade (a set of all leaves that share a common ancestor), we calculated the fraction of times it appears in the collected samples, which we call a *support* in this pa-

Features	Frequencies/Values
Consonant Inventories	95 Average 5 Moderately small
Vowel Quality Inventories	85 Average (5-6) 15 Small (2-4)
Syllable Structure	100 Moderately complex 0 Complex
Coding of Nominal Plurality	97 Plural suffix 2 Plural word 1 No plural 0 Plural clitic
Order of Numeral and Noun	61 Noun-Numeral 39 Numeral-Noun
Position of Case Affixes	61 No case affixes or adp. clitics 39 Case suffixes
Ord. of SOV	61 SOV 38 SVO 1 No dominant order
Ord. of Adposition and NP	91 Postpositions 9 Prepositions
Ord. of Adjective and Noun	87 Noun-Adjective 13 Adjective-Noun

Table 3: Some features of the world's ancestor with sample frequencies.

per. A tree was scored by the product of supports of all clades within it, and we created a tree that maximized the score. Each edge label shows the support of the corresponding clade. As indicated by generally low supports, the sample trees were very unstable. Some geographically distant groups of languages were clustered near the bottom. We partially attribute this to the *underspecificity* of linguistic typology: even if a pair of languages shares the same feature vector, they are not necessarily the same language. This problem might be eased by incorporating geospatial information into phylogenetic inference (Bouckaert et al., 2012).

Table 3 shows some features of the root. The reconstructed ancestor is moderate in phonological typology, uses suffixing in morphology and prefers the SOV word order. The inferred word order agrees with speculations given by previous studies (Maurits and Griffiths, 2014).

Figure 5 shows the histogram of variance parameters. Some latent components had smaller variances and thus were more stable with respect to evolutionary history. Figure 6 displays languages using the components with the two smallest variances. Unlike PCA plots, data concentrated at the edges.

Rank	Language	Classificatoin	Logprob.
1	(Japanese)	Japonic	76.8
2	Shuri	Japonic	-37.7
3	Khalkha	Altaic>Mongolic	-200.0
4	Lepcha	Sino-Tibetan>Tibeto-Burman	-201.9
5	Chuvash	Altaic>Turkic	-205.5
6	Deuri	Sino-Tibetan>Tibeto-Burman	-218.3
7	Urum	Altaic>Turkic	-218.6
8	Ordos	Altaic>Mongolic	-219.0
9	Uzbek	Altaic>Turkic	-219.6
10	Archi	N. Caucasian>E. Caucasian	-221.5
131	Korean	(isolate)	-265.7
493	Ainu	(isolate)	-409.9

Table 4: Modern languages ranked by the similarity to Japanese.

We used a geometric mean of p_{MOVE} of multiple samples to calculate how a modern language is similar to another. The case of Japanese is shown in Table 4. This ranked list is considerably different from that of disagreement rates of categorical vectors (Spearman's $\rho = 0.76$). When features' stability with respect to evolutionary history is considered, Japanese is less closer to Korean and Ainu than to some Tibeto-Burman languages south of the Himalayas. As the importance of these minor languages of Northeast India is recognized, the Sino-Tibetan tree might be drastically revised in the future (Blench and Post, 2013). The least similar languages include the Malayo-Polynesian and Nilo-Saharan languages.

7 Conclusion

In this paper, we proposed continuous space representations of linguistic typology and used them for phylogenetic inference. Feature dependencies are a major focus of linguistic typology, and typology data have occasionally been used for computational phylogenetics. To our knowledge, however, we are the first to integrate the two lines of research. In addition, the continuous space representations underlying interdependent discrete features are applicable to other data including phonological inventories (Moran et al., 2014).

We believe that typology provides important clues for long-term language change. The currently available database only contains modern languages, but we expect that data of some ancestral languages could greatly facilitate computational approaches to diachronic linguistics.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number 26730122.

References

Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Roger Blench and Mark W. Post. 2013. Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages. In Nathan Hill and Tom Owen-Smith, editors, *Trans-Himalayan Linguistics*, pages 71–104. De Gruyter.

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

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

William Croft, Tanmoy Bhattacharya, Dave Kleinschmidt, D. Eric Smith, and T. Florian Jaeger. 2011. Greenbergian universals, diachrony, and statistical analyses. *Linguistic Typology*, 15(2):433–453.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *ACL*, pages 65–72.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *HLT-NAACL*, pages 593–601.

Patricia Donegan and David Stampe. 2004. Rhythm and the synthetic drift of Munda. In Rajendra Singh, editor, *The Yearbook of South Asian Languages and Linguistics*, pages 3–36. Mouton de Gruyter.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *COLING*, pages 385–393.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Joseph H. Greenberg, editor. 1963. *Universals of language*. MIT Press.

Joseph H. Greenberg. 1978. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravesik, editors, *Universals of human language*, volume 1. Stanford University Press.

Simon J. Greenhill, Quentin D. Atkinson, Andrew Meade, and Russel D. Gray. 2010. The shape and tempo of language evolution. *Proc. of the Royal Society B*, 277(1693):2443–2450.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Shiro Hattori. 1999. *Nihongo no keito (The Genealogy of Japanese)*. Iwanami Shoten.

Katherine A. Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *ICML*, pages 297–304.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Julie Josse, Marie Chavent, Benot Liquet, and François Husson. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29(1):91–116.

Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. 2012. Efficient active algorithms for hierarchical clustering. In *ICML*, pages 887–894.

Walter J. Le Quesne. 1974. The uniquely evolved character concept and its cladistic application. *Systematic Biology*, 23(4):513–517.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2014. *Ethnologue: Languages of the World, 17th Edition*. SIL International. Online version: http://www.ethnologue.com.

Shuying Li, Dennis K. Pearl, and Hani Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 95(450):493–508.

Luke Maurits and Thomas L. Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *PNAS*, 111(37):13576–13581.

Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kevin P. Murphy. 2007. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia.

Johanna Nichols. 2011. Monogenesis or polygenesis: A single ancestral language for all humanity? In Maggie Tallerman and Kathleen R. Gibson, editors, *The Oxford Handbook of Language Evolution*, pages 558–572. Oxford Univ Press.

John Novembre and Matthew Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649.

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *COLING Posters*, pages 975–984.

Morris Swadesh. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proc. of American Philosophical Society*, 96:452–463.

Yee Whye Teh, Hal Daumé III, and Daniel Roy. 2008. Bayesian agglomerative clustering with coalescents. In *NIPS*, pages 1473–1480.

Nikolai Sergeevich Trubetzkoy. 1928. Proposition 16. In *Acts of the First International Congress of Linguists*, pages 17–18.

Alexander Vovin. 2005. The end of the Altaic controversy. *Central Asiatic Journal*, 49(1):71–132.