From Bag of Languages to Family Trees From Noisy Corpus

Taraka Rama and Anil Kumar Singh
Language Technologies Research Centre, IIIT, Hyderabad, India
taraka@students, anil@research@students.iiit.ac.in

Abstract

In this paper, we use corpus-based measures for constructing phylogenetic trees and try to address some questions about the validity of doing this and applicability to *linguistic areas* as against language families. We experiment with four corpus based distance measures for constructing phylogenetic trees. Three of these measures were earlier tried for estimating language distances. We use a fourth measure based on phonetic and orthographic feature n-grams. We compare the trees obtained using these measures and present our observations.

Keywords

Language distances, similarity measures, phylogenetic trees

1 Introduction

Establishing relationships among languages which have been in contact for a long time has been a topic of interest in historical linguistics [6]. However, this topic has been much less explored in the computational linguistics community. Most of the previous work is focused on reconstruction of phylogenetic trees for a particular language family using handcrafted word lists [12, 3, 2, 14] or using synthetic data [4].

In this paper we pose the following questions. What happens when we try to construct phylogenetic trees using inter-language distances in the context of a *linguistic area*[1]? Can the phylogenetic trees be used for evaluating the robustness of the inter-language distance measures and the meaningfulness of the distances? To our knowledge these questions have not been addressed previously. As Singh and Surana [18] showed, corpus based measures can be successfully used for comparative study of languages. Can these distances, estimated from a noisy corpus[2], meaningfully be used to construct phylogenetic trees? Can the information represented by the tree give meaningful interpretations about the languages involved? In this paper, we try to answer these questions. By using meaningful measures for estimating the distance between languages, we try to establish that the answers to these questions are affirmative. Overall, the contributions of the paper are the following a) use a new measure for estimating language distance b) present results of the experiments on constructing phylogenetic trees from corpus based word lists rather than handcrafted ones c) validate the hypothesis that India is a linguistic area [10].

The paper is organized as follows. Related work is discussed in Section 2. A brief discussion of various inter-language measures is given in Section 3. The experimental setup and the analysis of the results have been given in Section 4 and Section 5, respectively. We present a summary of our experiments, analysis of the results and future directions of the work in Section 6.

2 Related Work

In recent years, the methods developed in computational biology [13, 15, 11, 19] have been successfully adapted in computational linguistics for constructing the phylogeny[3]. All these methods are character based or distance based methods. The major disadvantage of these approaches is that they require handcrafted lists. Moreover, the methods inspired from glottochronology take a boolean matrix as input, which denotes the change in the state of the 'characters' (the 'characters' can be lexical, morphological or phonological) to infer the phylogenetic trees.

Ellison and Kirby [9] discuss establishing a probability distribution for every language through intra-lexical comparison using confusion probabilities. They use normalized edit distance to calculate the probabilities. Then the distance between every language pair is estimated as a distance between the probability distributions formed for individual languages. The distances (between languages) are estimated using KL-divergence and Rao's distance. The same measures are also used to find the level of cognacy between the words. The experiments are conducted on Dyen's [8] classical Indo-European dataset. The estimated distances are used for constructing a phylogenetic tree of the Indo-European languages.

Bouchard-Cote et al. [5], in a novel attempt, combine the advantages of classical comparative method and the corpus-based probabilistic models. The word forms are represented by phoneme sequences which un-

[1] The term *linguistic area* or *Sprachbund* [10] refers to a group of languages that have become similar in some way as a result of proximity and *language contact*, even if they belong to different families. The best known example is the Indian (or South Asian) linguistic area.

[2] By noisy corpus we mean a corpus that includes wrongly spelled words and spelling variations.

[3] Phylogeny is the (study of) evolutionary development and history of a species or higher taxonomic grouping of organisms. The term is now also used for other things such as tribes and languages. Phylogenetic trees represent this evolutionary development.

dergo stochastic edits along the branches of a phylogenetic tree. The robustness of the model is proved when it selects the linguistically attested phylogeny. The stochastic models successfully model the language change by using synchronic languages to reconstruct the word forms in Vulgar Latin and Classical Latin. Although it reconstructs the ancient word forms of the Romance Languages, a major disadvantage of this model is that some amount of data of the ancient word forms is required to train the model, which may not be available in many cases.

In another novel attempt, Singh and Surana [18] used corpus based simple measures to show that corpus can be used for comparative study of languages. They used both character n-gram distances and Surface Similarity [16] to identify the potential cognates[4], which in turn are being used to estimate the inter-language distance. Both diachronic and synchronic experiments are performed and the results very well attest to the linguistic facts. They also argued that there is a common orthographic as well as phonetic space for languages with a long history of contact which can be exploited for developing inter-language (rather than intra-language) measures, in contrast to the position taken by Ellison and Kirby [9]. Having followed this line of argument, we explain some corpus measures which we adopted from their work and also use a new measure which we call phonetic (and orthographic) feature n-gram based distance.

3 Inter-Language Measures

Such measures can be broadly divided into three categories. Character n-gram measures, cognate based measures and feature n-gram measures. The following sections describe each measure in more detail. One important point that can be mentioned here is that all the languages we experimented on use Brahmi origin scripts, which have almost one-to-one correspondence between letters and phonemes. Moreover, these scripts are similar in a lot of ways, especially the fact that the alphabets used by them can be seen as subsets of the same abstract alphabet, although the letters may have different shapes so that to a lay person the scripts seem very different. In fact, there is a 'super encoding' or 'meta encoding' called ISCII that can be used to represent this common alphabet. The letters of this common alphbet can be approximately treated like phonemes for computational purposes. For languages which do not use such scripts, we will first have to convert the text into a phonetic notation to be able to use the methods described below, except perhaps the first one.

3.1 Symmetric Cross Entropy (SCE)

The first measure is purely a letter n-gram based measure similar to the one used by Singh [17] for language and encoding identification. Note that since letters in Brahmi origin scripts can almost be treated like phonemes, we could call this method a phoneme n-gram based measure. To calculate the distance, letter 5-gram models are prepared from the corpora of the languages to be compared. Then the n-grams of all sizes (unigrams, bigrams, etc.) are combined and sorted according to their probability in descending order. Only the top N n-grams are retained and the rest are pruned. This is based on the results obtained by Cavnar [7] and validated by Singh, which show that the top N (300 according to Cavnar) n-grams have a high correlation with the identity of the language. At this stage there are two probability distributions which can be compared by a measure of distributional similarity. The measure used here is symmetric cross entropy:

$$d_{sce} = \sum_{g_l = g_m} (p(g_l)\ log\ q(g_m) + q(g_m)\ log\ p(g_l)) \quad (1)$$

where p and q are the probability distributions for the two languages and g_l and g_m are n-grams in languages l and m, respectively. The probabilities of bigrams and larger n-grams are relative frequencies over a single distribution consisting of n-grams of all sizes up to 5 (the 'order' of the n-gram model), not conditional probabilities, as in standard n-gram models for calculating sequence probabilities.

The disadvantage of this measure is that it does not use any linguistic (e.g., phonetic) information, but the advantage is that it can easily measure the similarity of distributions of n-grams. Such measures have proved to be very effective in automatically identifying languages of text, with accuracies nearing 100% for fairly small amounts of training and test data [1, 17].

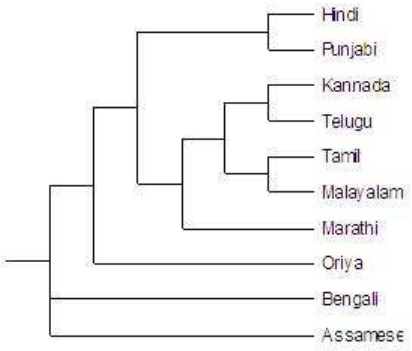

Fig. 1: *Phylogenetic tree using SCE*

3.2 Measures based on Cognate Identification

The other two measures are based on potential cognates, i.e., words of similar form. Both of them use an algorithm for identification of potential cognates. Many such algorithms have been proposed. For identifying cognates, Singh and Surana [18] used the Computational Phonetic Model of Scripts or CPMS [16]. This model takes into account the characteristics of Brahmi origin scripts and calculates Surface Similarity.

[4] Potential cognates are words of different languages which are similar in form and therefore are likely to be cognates. They might include some 'false friends', i.e., words which are not etymologically inherited. It is worthwhile to experiment (using statistical techniques) on potential cognates, even without removing the 'false friends' because a large percentage of them are actually cognates in the linguistic sense.

It consists of a model of alphabet that represents the common alphabet for Brahmi origin scripts, a model of phonology that maps the letters (which are, for the most part, phonemes) to phonetic and orthographic features, a Stepped Distance Function (SDF) that calculates the phonetic and orthographic similarity of two letters and a dynamic programming (DP) algorithm that calculates the Surface Similarity of two words or strings. The CPMS was adapted by Singh and Surana for identifying the potential cognates.

In general, the distance between two strings can be defined as:

$$c_{lm} = f_p(w_l, w_m) \tag{2}$$

where f_p is the function (implemented as a DP alignment algorithm) which calculates Surface Similarity using the CPMS based cost between the word w_l of language l and the word w_m of language m.

Those word pairs are identified as cognates which have the least cost.

3.2.1 Cognate Coverage Distance (CCD)

The second measure used is a corpus based estimate of the coverage of cognates across two languages. Cognate coverage is defined ideally as the number of words (from the vocabularies of the two languages) which are of the same origin, but which is approximately estimated by identifying words of similar form (potential cognates). The decision about whether two words are cognates or not is made on the basis of Surface Similarity of the two words as described in the previous section. Non-parallel corpora of the two languages are used for identifying the cognates.

The normalized distance between two languages is defined as:

$$t'_{lm} = 1 - \frac{t_{lm}}{max(t)} \tag{3}$$

where t_{lm} and t_{ml} are the number of (potential) cognates found when comparing from language l to m and from language m to l, respectively.

Since the CPMS based measure of Surface Similarity is asymmetric, the average number of unidirectional cognates is calculated:

$$d^{ccd} = \frac{t'_{lm} + t'_{ml}}{2} \tag{4}$$

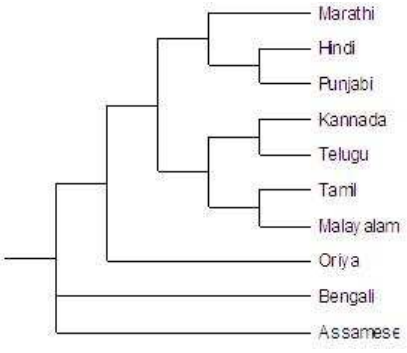

Fig. 2: *Phylogenetic tree using CCD*

3.2.2 Phonetic Distance of Cognates (PDC)

Simply finding the coverage of cognates may indicate the distance between two languages, but a measure based solely on this information does not take into account the variation between the cognates themselves. To include this variation into the estimate of distance, Singh and Surana [18] used another measure based on the sum of the CPMS based cost of n cognates found between two languages:

$$C^{pdc}_{lm} = \sum_{i=0}^{n} c_{lm} \tag{5}$$

where n is the minimum of t_{lm} for all the language pairs compared.

The normalized distance can be defined as:

$$C'_{lm} = \frac{C^{pdc}_{lm}}{max(C^{pdc})} \tag{6}$$

A symmetric version of this cost is then calculated:

$$d_{pdc} = \frac{C'_{lm} + C'_{ml}}{2} \tag{7}$$

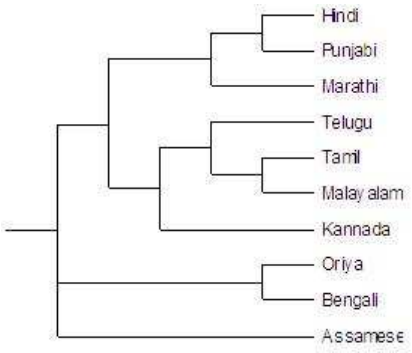

Fig. 3: *Phylogenetic tree using PDC*

3.3 Feature N-Grams (FNG)

The idea in using this measure is that the way phonemes occur together matters less than the way the phonetic features occur together because phonemes themselves are defined in terms of the features. Therefore, it makes more sense to a have measure directly in terms of phonetic features. But since we are experimenting directly with corpus data (without any phonetic transcription) using the CPMS [16], we also include some orthographic features as given in the CPMS implementation. The letter to feature mapping that we use comes from the CPMS. Basically, each word is converted into a set of sequences of feature-value pairs such that any feature can follow any feature, which means that the number of sequences for a word of length l_w is less than or equal to $(N_f \times N_v)^{l_w}$, where N_f is the number of possible features and N_v is the number of possible values. We create sequences of feature-value pairs for all the words and from this 'corpus' of feature-value pair sequences we build the feature n-gram model.

The formula for calculating distributional similarity based on these phonetic and orthographic features is the same (SCE) as given in equation 1, except that the distribution in this case is made up of features rather than letters. Note that since we do not assume the features to be independent, any feature can follow any other feature in a feature n-gram. All the permutations are computed before the feature n-gram model is pruned to keep only the top N feature n-grams. The order of the n-gram model is kept as 3, i.e., trigrams.

The feature n-grams are computed as follows. For a given word, each letter is first converted into a vector consisting of the feature-value pairs which are mapped to it by the CPMS. Then, from the sequence of vectors of features, all possible sequences of features up to the length 3 (the order of the n-gram model) are computed. All these sequences of features (feature n-grams) are added to the n-gram model. Finally the model is pruned as mentioned above. We expected this measure to work better because it works at a higher level of abstraction and is more linguistically valid.

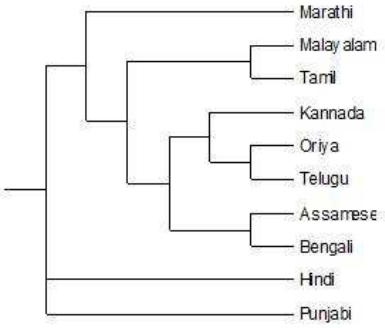

Fig. 4: *Phylogenetic tree using feature n-grams*

4 Experimental Setup

Although the languages we selected belong to two different language families, there are a lot of similarities among them which allow us to choose them for our experiments [10]. The corpora used for our experiments are all part of the CIIL multilingual corpus. The experiments were conducted using word lists prepared from the raw corpus for every language. No morph analyzer or stemmer has been applied to the words. Initially the word types with their frequencies are extracted from the corpus. Then the word types are sorted based on their corresponding frequency. Only the top N_w of these word types are retained. This is done with the aim of including as much of the core vocabulary as possible for comparing the languages[5]. For using cognate based measures for estimation of language distance, cognates are extracted from the word lists between these languages. For feature n-gram measures, the feature n-gram models are prepared as explained in Section 3.

[5] For our experiments we fixed N_w at 50,000. This number is different from N, the number of top n-grams that are retained after pruning the n-gram model.

We calculate the distance between every pair of languages available. We compare the results between all the four measures discussed above by constructing trees using these measures. The trees are constructed using the NEIGHBOR program in the PHYLIP package[6]. The NEIGHBOR programs provides two distance-based tree construction algorithms: Neighbour Joining and UPGMA. For our experiments we used Neighbour Joining as it does not assume a constant rate of evolution and it produces unrooted trees unlike UPGMA which assumes constant rate of evolution (the length of the leaves from the root of the tree is same across all the leaves) and produces rooted trees. We do not do any outgrouping as outgrouping makes sense only when all the languages belong to a single family.

	BN	**HI**	**KN**	**ML**	**MR**	**OR**	**PA**	**TA**	**TE**
AS	0.02	0.39	0.71	0.86	0.61	0.20	0.61	0.93	0.73
	0.12	0.25	0.39	0.61	0.45	0.11	0.58	0.95	0.46
	0.05	0.30	0.51	0.50	0.43	0.18	0.42	0.70	0.64
	0.02	0.06	0.07	0.12	0.09	0.05	0.09	0.13	0.05
BN		0.32	0.68	0.86	0.57	0.07	0.56	0.96	0.70
		0.29	0.42	0.64	0.42	0.05	0.56	0.90	0.50
		0.29	0.47	0.45	0.43	0.14	0.42	0.74	0.43
		0.06	0.07	0.13	0.08	0.04	0.09	0.11	0.02
HI			0.61	0.81	0.42	0.40	0.20	0.93	0.61
			0.17	0.56	0.16	0.27	0.16	0.87	0.38
			0.43	0.46	0.16	0.33	0.20	0.74	0.34
			0.09	0.09	0.06	0.08	0.03	0.15	0.13
KN				0.77	0.68	0.75	0.73	0.88	0.53
				0.45	0.17	0.31	0.50	0.82	0.25
				0.18	0.38	0.52	0.58	0.42	0.09
				0.10	0.09	0.02	0.08	0.10	0.03
ML					0.89	0.88	0.88	0.62	0.72
					0.65	0.59	0.77	0.56	0.31
					0.42	0.53	0.55	0.07	0.19
					0.13	0.13	0.11	0.07	0.15
MR						0.64	0.52	0.95	0.68
						0.40	0.37	0.94	0.46
						0.34	0.39	0.60	0.30
						0.08	0.06	0.13	0.09
OR							0.63	0.98	0.74
							0.45	0.89	0.44
							0.65	0.83	0.64
							0.07	0.10	0.00
PA								0.90	0.71
								0.90	0.59
								0.92	0.48
								0.14	0.07
TA									0.85
									0.81
									0.39
									0.08

AS: Assamese, **BN:** Bengali, **HI:** Hindi, **KN:** Kannada
ML: Malayalam, **MR:** Marathi, **OR:** Oriya,
PA: Punjabi, **TA:** Tamil, **TE:** Telugu

Table 1: *Inter-language comparison among ten major South Asian languages using four corpus based measures. The values have been normalized and scaled to be somewhat comparable. Each cell contains four values: by CCD, PDC, SCE and FNG.*

[6] `http://evolution.genetics.washington.edu/phylip/phylip.html`

5 Analysis of Results

Table 1 shows the results obtained for the four distance measures. Figures 1 to 4 show the trees obtained using all the above measures. There are three subgroupings of the languages which are clearly visible in all the trees. Namely, Northern Indo-Aryan (Hindi and Punjabi), Eastern Indo-Aryan (Assamese, Bengali and Oriya) and Dravidian languages (Tamil, Kannada, Malayalam and Telugu). There are clearly some similarities in the trees which are generated by all the methods. All the methods group Hindi and Punjabi, Tamil and Malayalam together. CCD gives the normalized measure of the number of cognates between every language pair. In the case of CCD tree, although Bengali and Assamese are grouped together, Oriya is placed incorrectly, which is correctly placed in the case of feature n-grams.

Oriya is incorrectly grouped with Bengali in the case of PDC tree. The reason can be because of the huge number of shared words which cause a lower phonetic distance between the languages. Kannada and Telugu are not grouped together in the case of PDC. Marathi is either classified with Northern Indo-Aryan languages or with Dravidian languages. It is grouped with Indo-Aryan languages in the case of cognate distance measures and grouped with Dravidian languages in the other cases. The reason for grouping it with Dravidian languages is the influence of Dravidian languages due to long history of contact.

The distance of a terminal node from its parent gives very important information[7]. For example, Tamil is always at a greater distance from its parent node, although grouped with Malayalam, compared to other languages. Especially in the case of feature n-grams and SCE, the distance is very evident. The reason for this is the lower number of 'characters' (elements from which n-grams are made) when compared to other languages in the case of SCE. In the case of feature n-grams, the lack of phonemic distinction in writing between voiced and unvoiced sounds for Tamil decreases the number of shared feature n-grams. Moreover, the number of borrowings from Indo-Aryan Languages are comparatively less in the case of Tamil.

6 Conclusion and Future Work

In this paper we discussed the possibility of using corpus based measures for constructing phylogenetic trees. Four corpus based measures were used for the construction of phylogenetic trees. Out of these measures, the second, the third and the fourth measure are linguistically well grounded measure. We considered the differences between each tree and tried to explain the reasons for the anomalies in the tree structure. We have shown that by using noisy corpus and simple but linguistically well founded measures, we can very nearly achieve the desired family tree. These measures can be very useful for languages which do not have linguistically hand-crafted lists. The experiments also demonstrate that the technique can be applicable even to *linguistic areas*, not just language families.

[7] The trees in the figures are not scaled, but the distances are given in the table.

Acknowledgment

We would like to acknowledge the valuable suggestions made by Sudheer Kolachina from LTRC, IIIT, Hyderabad. We are also thankful to the reviewers for their constructive comments which we tried to take into account as far as possible.

References

[1] G. Adams and P. Resnik. A Language Identification Application Built on the Java Client/Server Platform. *From Research to Commercial Applications: Making NLP Work in Practice*, pages 43–47, 1997.

[2] Q. Atkinson and R. Gray. How old is the Indo-European language family? Progress or more moths to the flame. *Phylogenetic Methods and the Prehistory of Languages (Forster P, Renfrew C, eds)*, pages 91–109, 2006.

[3] Q. Atkinson, G. Nicholls, D. Welch, and R. Gray. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219, 2005.

[4] F. Barbançon, T. Warnow, S. Evans, D. Ringe, and L. Nakhleh. An experimental study comparing linguistic phylogenetic reconstruction methods. Technical report, Technical Report 732, Department of Statistics, University of California, Berkeley, 2007.

[5] A. Bouchard-Cote, P. Liang, T. Griffiths, and D. Klein. A probabilistic approach to language change. NIPS, 2000.

[6] L. Campbell. *Historical linguistics: an introduction*. MIT Press, 2004.

[7] W. Cavnar and J. Trenkle. N-gram-based text categorization. *Ann Arbor MI*, 48113:4001, 1994.

[8] I. Dyen, J. Kruskal, and P. Black. An Indoeuropean classification: a lexicostatistical experiment. Amer Philosophical Society, 1992.

[9] T. Ellison and S. Kirby. Measuring language divergence by intra-lexical comparison. In *Proceedings of the 44th annual meeting of the ACL*, pages 273–280. Association for Computational Linguistics Morristown, NJ, USA, 2006.

[10] M. Emeneau. India as a Lingustic Area. *Language*, pages 3–16, 1956.

[11] J. Felsenstein. Inferring Phylogenies. Sunderland, MA. *Sinauer Press. Chapters*, 1(7):11, 2003.

[12] R. Gray and Q. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Earth Planet. Sci*, 23:41–63, 1995.

[13] J. Huelsenbeck, F. Ronquist, R. Nielsen, and J. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314, 2001.

[14] L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420, 2005.

[15] N. Saitou. The neighbor-joining method: a new method for reconstructing phylogenetic trees, 1987.

[16] A. K. Singh. A computational phonetic model for indian language scripts. In *Proceedings of the Constraints on Spelling Changes: Fifth International Workshop on Writing Systems*, Nijmegen, The Netherlands, 2006.

[17] A. K. Singh. Study of some distance measures for language and encoding identification. In *Proceeding of ACL 2006 Workshop on Linguistic Distances*, Sydney, Australia, 2006.

[18] A. K. Singh and H. Surana. Can corpus based measures be used for comparative study of languages. In *Proceedings of the ACL Workshop Computing and Historical Phonology*, Prague, Czech Republic, 2007.

[19] D. Swofford, G. Olsen, P. Waddell, and D. Hillis. Phylogenetic inference. *Molecular systematics*, 2:407–514, 1996.