Harnessing Deep Cross-lingual Word Embeddings to Infer Accurate Phylogenetic Trees

Yashasvi Mantha[5], Diptesh Kanojia[1,2,3], Abhijeet Dubey[4], Pushpak Bhattacharyya[1], and Malhar Kulkarni[1]

[1]IIT Bombay, [2]IITB-Monash Research Academy, [3]Monash University, [4]Apple, [5]GITAM University

{yashasvimantha, dipteshkanojia, dubey.abhijeet09}@gmail.com

Introduction

- India is one of the most religiously and ethnically diverse nations in the world which also makes it home to various Languages.
- With so many languages in the country, it is necessary for us to understand the relatedness of various Indian languages.
- In this paper we try to answer questions about the same.
- We will be focusing on construction of Phylogenetic trees that can reveal various details of the closeness.
- In this paper, we utilize fourteen linked Indian Wordnets to create inter-language distances using our novel approach to compute 'language distances'.
- The traditional methods for the construction of various phylogenetic trees do not take the scemantics of the word. Hence also not taking the meaning of the word while calculating the distance matrix.
- We use two different approaches to construct the language distance matrix. Here by distance matrix we mean the closness of each language pair. Since we are utilizing 14 diffrent indian languages, our distance matrix would consist of 196 entries.
- Each language's corpus ranges for about two Lakh to about 15 thousand lines.

Data

- Primarily we use the parallel corpus and WordNet from the IndoWordNet (Bhattacharyya, 2017) dataset for the experiments. We use the datasets of 14 different Indian Languages detailed in the table below:

Hindi(41K)	Marathi(54K)	Bengali(100K)
Assamese(15K)	Kannada(22K)	Malayalam(39K)
Gujarati(103K)	Oriya(35K)	Konkani(32K)
Nepali(200K)	Telugu(36K)	Sanskrit(150K)
Tamil(36K)	Punjabi(36K)	

- WordNets are organised in a thesaurus way (in a sense order). Which means every ID across a family of WordNets has the same context.
- Here we use only monolingual corpus from which we make crosslingual or multilingual corpus.
- The WordNets we are utilising have 5 parts. Each part is delimited with a semi-colon(";"). The format is described below: [1]

$$(ID\ ;\ Words\ ;\ Gloss\ ;\ Definition\ ;\ POS) \tag{1}$$

- Since the Indian language we experiment upon have diffrent scripts, we had to convert all the languages to a common script (Devanagari) (Anoop Kunchukuttan, 2013).

Calculation of the Distance Matrix

- The distance matrix reptesents the dis-similarities of each label pair (here language pair) which we will calculate.
- There will be two distance matrix that we will calculate (one for baseline and one for our novel approach). These two matrices were calculated using completely diffrently approaches.
- The baseline approach uses weighted lexical similarity measure to calculate the distance. The average of word-pair distances provides us 'synset distance' and further averaging of parallel synset distances provides us a baseline inter-language distance.
- We use word embeddings (Mikolov et al., 2013) to effectiently represent the words. These word embeddings are then subjected to various computational processes to find the similarity.
- Our novel approach computes the angular cosine distance (Cer et al., 2018) between all word pairs belonging to the same synset in the common embedding space shared by two languages.
- Both Monolingual and crosslingual word embeddings were calculated using fastText and Muse (Joulin et al., 2016).
- These word embeddings were then subjected to angular cosine distance to calculate the language pair distance.
- While computing monolingual word embeddings a dimention size of 50 was found effective.
- The use of word embeddings reduced the size of data exportially as compared to one-hot encoding.
- The same language pair was discarded for computation and replaced with the ideal case of 0.
- Some language pair distances of our baseline and novel approach are listed below:

Baseline approach

Language Pair	Distance
as – bn	0.7885
ta - te	0.8973
pa – hi	0.7993

Novel Approach

Language Pair	Distance
as – bn	0.4299
ta - te	0.4172
pa – hi	0.4163

Construction of Phylogenetic Trees

- Phylogenetic trees can be constructed using various computational phylogenetic methods. Here, we majorly use the distance based approaches which requires a distance matrix that contains the distance between each label.
- The distance based methods use a distance matrix to construct phylogenetics trees.
- Here we effectively use the UPGMA or Unweighted Pair Group method with arithmetic Mean (Sokal, 1958) where the basic idea is the combine the two nearest clusters into a higher node removing and centering the initial nodes selected.
- The distance between ant two clusturs is given by equation 2.

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y) \tag{2}$$

- We used Fionn Murtagh's algorithm (Day and Edelsbrunner, 1984) for k-dimensional data that has a time complexity of $O(n^2)$ for constant k.
- It is worth noting that a phylogenetic tree does not necessarly have all the nodes labled. But at the same time it is necessary for phylogenetic trees to have labels for leaf nodes (i.e. nodes that do not have any children).
- For the representation of trees we used the newick format which is a mathamatical way of representation of trees.

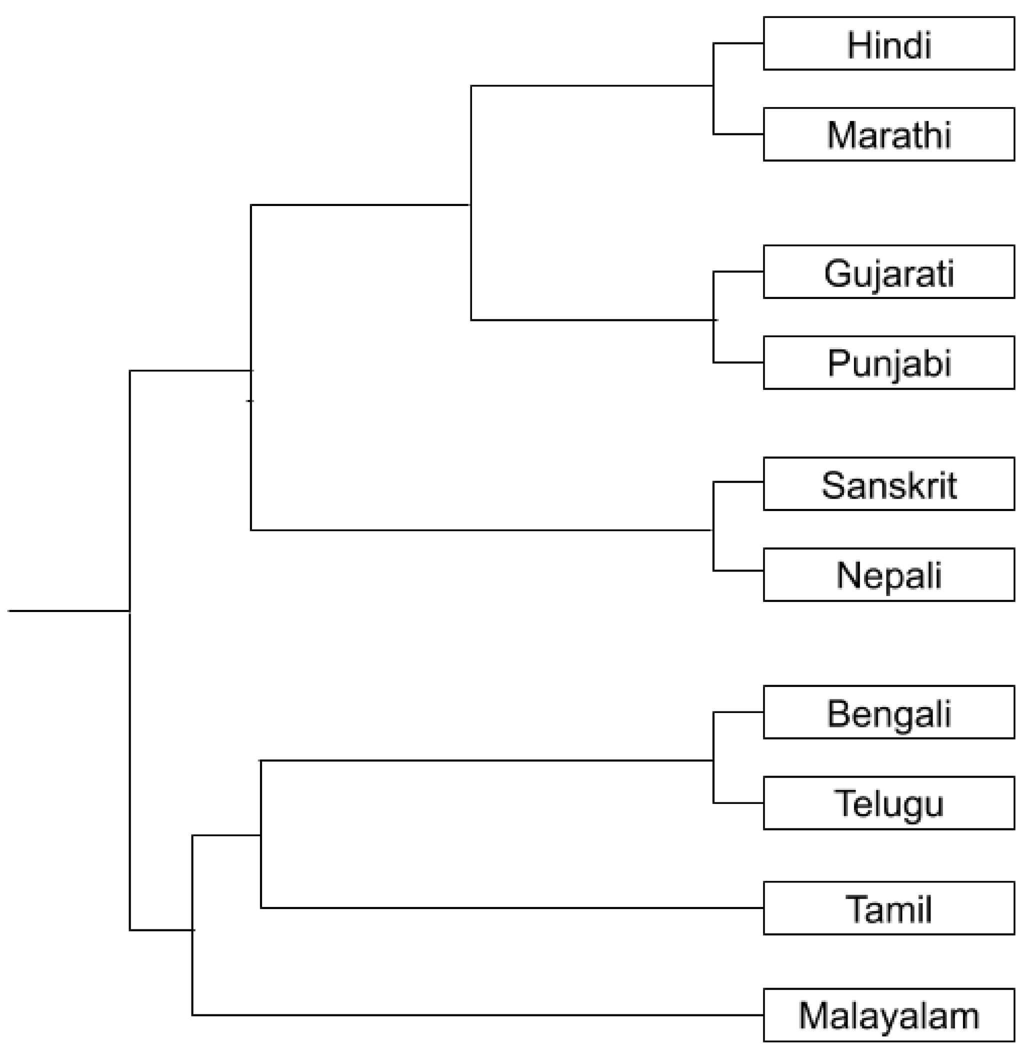

Figure 1: *The outputed tree from baseline approach*

- It is worth noting that a phylogenetic tree does not necessarly have all the nodes labled. But at the same time it is necessary for phylogenetic trees to have labels for leaf nodes (i.e. nodes that do not have any children).
- For the representation of trees we used the newick format which is a mathamatical way of representation of trees.

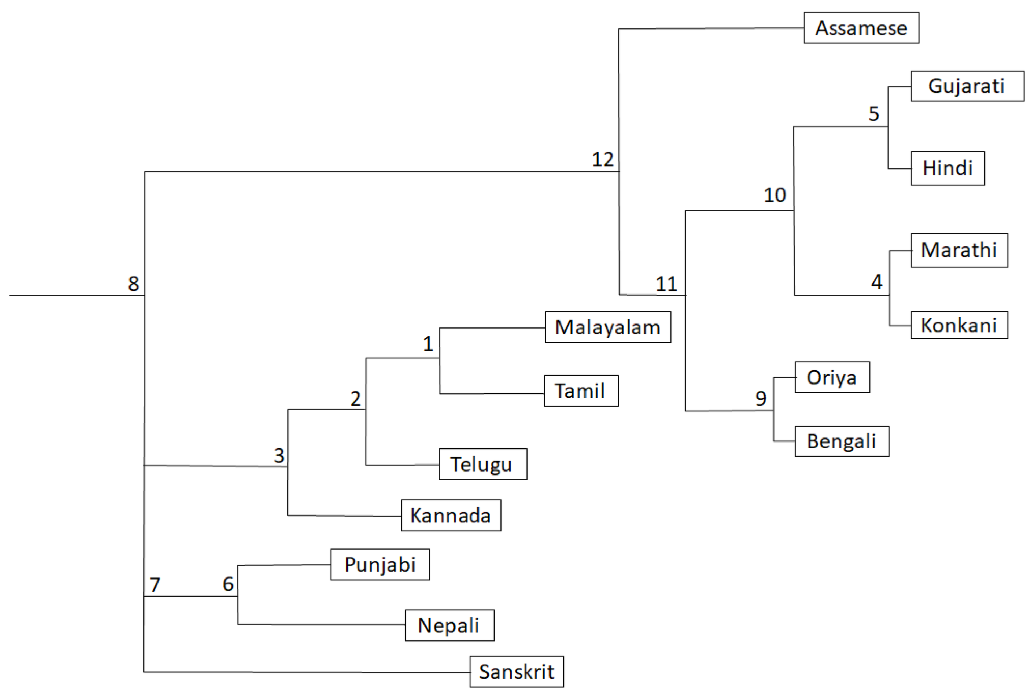

Figure 2: *The outputed tree form out Novel approach*

Conclusion and Future Work

- In this paper, we come up with a methodology for the construction of phylogenetics of 14 diffrent Indian Languages.
- We propose the word embeddings methodology for the construction and see that our novel approach clearly performs better than that of the traditional edit-distance approach.
- We train deep cross-lingual word embeddings for every language pair and use angular cosine distance to compute distance matrices.
- We also hypothise that adding potential cognate data would result in better trees.
- We want to add other Indian languages and increase the corpora size along with different cross-lingual embeddings to further substantiate our claim.
- We also think that the accuracy of the deep cross-lingual word embeddings can be substantually improved.
- The word vectors that was resulted form fastText (Joulin et al., 2016) can also effectively be improved.

References

Anoop Kunchukuttan (2013). Indic nlp library. `https://github.com/anoopkunchukuttan/indic_nlp_library`.

Bhattacharyya, P. (2017). Indowordnet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Day, W. H. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438.

[1] *We acknowledge and thank the IndoWordnet Dataset creators. †diptesh@cse.iitb.ac.in