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

Yashasvi Mantha, Diptesh Kanojia[†], Abhijeet Dubey, Pushpak Bhattacharyya, and Malhar Kulkarni[*]
Center for Indian Language Technology, Indian Institute of Technology, Bombay

ABSTRACT

Establishing language relatedness by inferring phylogenetic trees has been a topic of interest in the area of diachronic linguistics. However, existing methods face meaning conflation deficiency due to the usage of lexical similarity-based measures. In this paper, we utilize fourteen linked Indian Wordnets to create inter-language distances using our novel approach to compute 'language distances'. Our pilot study uses deep cross-lingual word embeddings to compute inter-language distances and provide an effective distance matrix to infer phylogenetic trees. We also develop a baseline method using lexical similarity-based metrics for comparison and identify that our approach produces better phylogenetic trees which club related languages closer when compared to the baseline approach.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Natural language processing*; • **Information systems** → Information retrieval.

KEYWORDS

phylogenetic trees, historical linguistics, cross-lingual embeddings

1 INTRODUCTION AND MOTIVATION

Under the purview of diachronic linguistics, establishing relationships among languages which have been in contact for a long time has been a topic of interest [7]. Previous literature focuses on the reconstruction of phylogenetic trees for a language family using manually curated word lists [1, 2, 10, 12] or using synthetic data [3]. Word lists based measures can calculate the inter-language distance, but they use feature n-grams and cognates based methods which do not take into account the semantics of a word. Inspired by the recent trend in the usage of embeddings for estimating the semantics of a word, this paper proposes to use deep cross-lingual word embeddings (CWE) [8] to find the inter-language distances based on 'concepts' or 'synsets' [4, 9]. We hypothesize 'synset distance' based on wordnet data and utilize it to calculate 'inter-language distance'. We compute a distance matrix containing inter-language distances and utilize this distance matrix build phylogenetic trees.

2 DATASET AND APPROACHES

We build our dataset by Unicode offsetting the IndoWordnet data and investigate language pairs for Indian languages namely Hindi (Hi), Marathi (Mr), Konkani (Ko), Gujarati (Gu), Bengali (Bn), Oriya (Or), Assamese (As), Punjabi (Pa), Sanskrit (Sa), Tamil (Ta), Telugu(Te), Malayam (Ml), Kannada (Kn), and Nepali (Ne). For building CWE using MUSE [8], we use sub-word information enriched embeddings created using fastText [5]. Our corpora size ranges from ~25K lines (Kn) to ~48124K lines (Hi). We use two different approaches to construct the language distance matrix required by UPGMA [11] method, as detailed below. As a baseline approach, we use a weighted lexical similarity measure to calculate the distance matrix. *The average of word-pair distances provides us 'synset distance' and further averaging of parallel synset distances provides us a baseline inter-language distance.* **Our novel approach computes the angular cosine distance [6] between all word pairs belonging to the same synset in the common embedding space shared by two languages.** Thus, the average over the word-pair distances, and further 'synset distances' provides us with a more effective 'inter-language distance'. We use the UPGMA method to construct phylogenetic tree (Figure 1) of all the language pairs.

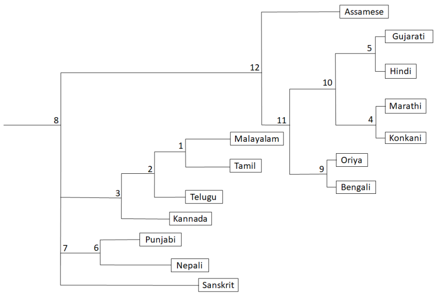

Figure 1: Resultant Tree Using Our Novel Approach

3 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel approach for calculating the distance matrix used to create phylogenetic trees through UPGMA. We hypothesize 'synset distance' from linked Wordnet data and successfully use it to calculate 'inter-language distances' for fourteen Indian languages. We train deep cross-lingual word embeddings for every language pair and use angular cosine distance to compute distance matrices. We produce matrices using a baseline approach and our novel approach and generate trees. We find that trees from our approach depict closeness in the languages better than the baseline and release our code and dataset[1]. In future, we would like to include more language families and increase the corpora size along with different cross-lingual embeddings to further substantiate our claim.

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

[1] http://www.cfilt.iitb.ac.in/typological

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371210

REFERENCES

[1] Quentin Atkinson, Geoff Nicholls, David Welch, and Russell Gray. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103, 2 (2005), 193–219.

[2] Quentin D Atkinson and Russell D Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame. *Phylogenetic methods and the prehistory of languages* 91 (2006), 109.

[3] François Barbançon, Steven N Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30, 2 (2013), 143–170.

[4] Pushpak Bhattacharyya. 2017. IndoWordNet. In *The WordNet in Indian Languages*. Springer, 1–18.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[7] F Chevillet. 2000. L. CAMPBELL.-" Historical Linguistics: An Introduction"(Book Review). *Études Anglaises* 53, 1 (2000), 59.

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017).

[9] Christiane Fellbaum. 2012. WordNet. *The Encyclopedia of Applied Linguistics* (2012).

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

[11] Ilan Gronau and Shlomo Moran. 2007. Optimal implementations of UPGMA and other common clustering algorithms. *Inform. Process. Lett.* 104, 6 (2007), 205–210.

[12] Luay Nakhleh, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* (2005), 382–420.