Cognate Identification to Improve Phylogenetics for Indian Languages

Diptesh Kanojia[1,2,3], Pushpak Bhattacharyya[2], Malhar Kulkarni[2], Gholamreza Haffari[3]
IITB-Monash Research Academy, IIT Bombay, Monash University

Cognate Identification and Phylogenetics: An Introduction

Cognates are words derived from the same origin into one or more languages *i.e.,* they have the same etymological origin.

Cognates are present in multiple variants of the same text across different languages.

Computational Phylogenetics uses algorithms and techniques to analyze these variants and infer phylogenetic trees for a hypothesized accurate representation based on the output of the computational algorithm used.

The study of cognates plays a crucial role in applying comparative approaches for historical linguistics, in particular, solving language relatedness and tracking the interaction and evolvement of multiple languages over time.

Motivation

Cognate detection helps phylogenetic inference by helping isolate diachronic sound changes and thus detect the words of a common origin.

Achieving good performance on automatic cognate identification can also benefit machine translation when dealing with two languages that share a certain quantity of cognates, as cognates are usually translations and serve as anchors when aligning.

A cognate instance in Indian languages is given as the word group: *putra* (Sanskrit), *putra* (Hindi), *putra* (Marathi) and *puttar* (Punjabi), all of which mean the word “Son”.

Related Work

Previous studies on cognate identification do not study Indian languages.

String similarity based methods are used as the baseline in the cognate detection papers (Melamed, 1999). We have also incorporated XDice (Brew *et. al., 1996)*, which is a set based similarity measure.

Research in automatic cognate identification using phonetic aspects involve computation of similarity by decomposing phonetically transcribed words (Kondrak, 2000), acoustic models (Mielke *et. al.*, 2012), phonetic encodings (Rama *et. al.,* 2015), aligned segments of transcribed phonemes (List *et. al.*, 2012).

IndoWordNet (Bhattacharyya, 2010) is a linked wordnet comprising of wordnets for major Indian languages.

Dataset Creation

We create the dataset by extracting word list for Hindi, Marathi, Sanskrit , and Punjabi WordNets. We transliterate the words in the Punjabi wordlist using Google Transliterate.

We use the unique words from the wordlist extracted from all the individual wordnet databases publicly available[1], but maintain them within the ID space.

We extract 15000 unique words from all the Wordnets and create wordlists aligned as per the synset ID.

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

Methodology

We use the baseline measure XDice and string similarity based measures to first prepare cognate sets from every individual language pair.

We construct more cognate sets with the use of Orthographic cognate detection methods such as alignment of substrings.

We use the phonetic aspects of the words decomposing them phonetically and aligning them according to phonemes.

We use string similarity measures and use the threshold value of 0.75 arrived at by empirical measures. We use Jaccard, XDice and TF-IDF are used to validate our cognate sets.

Statistics & Results

	Hindi - Punjabi	Hindi - Marathi	Hindi - Sanskrit
True Cognates	**497**	**621**	**378**
False Cognates	301	284	211
Total Detected	798	905	589

Textual History Tool

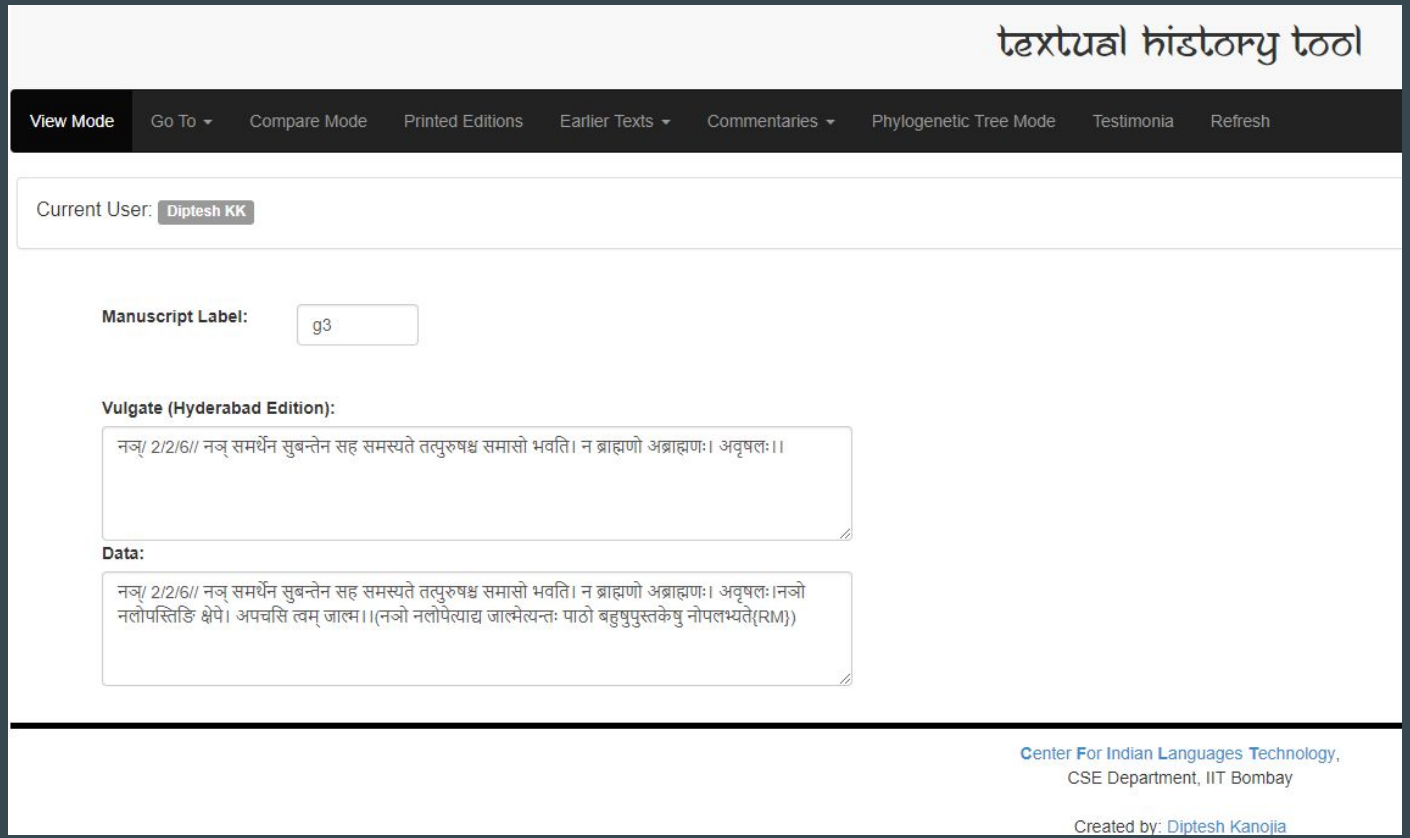
We also implement our work with Textual History Tool[1] and verify the impact of our cognate sets on the creation of phylogenetic trees.

The tool was created to facilitate the input of manuscript data and its variants digitally, and facilitates the creation of phylogenetic trees based on Maximum Likelihood and String similarity based measures.

We verify that inducing cognate words along with the manuscript variants indeed helps in the creation of better phylogenetic trees.

[1]http://www.cfilt.iitb.ac.in/~yogyata/4/admin/login.php

Textual History Tool: View Mode



Insights

While arriving at this value, we observed that we could easily form pairs of cognate words which are *Tatsama* words.

On the other hand, *Tadbhava* words were hardly detected among the cognate words unless phonetic methodologies were not used.

This poses a new challenge as *Tadbhava* word form a large set of cognate words among the Indian languages.

This can also be verified intuitively as the former retain their orthographic form and are easy to detect via the string similarity measure and the orthographic measure but the latter need phonetic measures.

Conclusion and Future Work

We describe our work on cognate detection for Indian language pairs Sanskrit - Hindi, Sanskrit - Marathi, and Sanskrit - Punjabi.

In the phylogenetic tree creation mode, we verify that cognate sets help in better phylogenetic tree creation.

We also release this cognate set dataset publicly. In this pilot study, we create cognate categorization and the nuances of cognate detection for Indian languages.

In future, we aim to expand our dataset to multiple Indian languages as wordlists in their root form are available publicly via the Indowordnet website.

References

Bhattacharyya, P. (2010). Indowordnet. In In Proc. of LREC-10. Citeseer.

Brew, C., McKelvie, D., et al. (1996). Word-pair extraction for lexicography. In Proceedings of the 2nd International Conference on New Methods in Language Processing, pages 45–55.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pages 288–295. Association for Computational Linguistics.

List, J.-M. (2012). Lexstat: Automatic detection of cognates in multilingual wordlists. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 117–125. Association for Computational Linguistics.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. Computational Linguistics, 25(1):107–130.

Mielke, M. M., Roberts, R. O., Savica, R., Cha, R., Drubach, D. I., Christianson, T., Pankratz, V. S., Geda, Y. E., Machulda, M. M., Ivnik, R. J., et al. (2012). Assessing the temporal relationship between cognition and gait: slow gait predicts cognitive decline in the mayo clinic study of aging. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 68(8):929–937.

Rama, T., Borin, L., Mikros, G., and Macutek, J. (2015). Comparative evaluation of string similarity measures for automatic language classification.

Thank You

Questions?