Utilizing Word Embeddings based Features for Phylogenetic Tree Generation of Sanskrit Texts

Diptesh Kanojia, Abhijeet Dubey, Malhar Kulkarni, Pushpak Bhattacharyya, and Reza Haffari
IITB-Monash Research Academy
IIT Bombay
Monash University
diptesh@cse.iitb.ac.in

Contribution

Overall contribution of this work:

- Break the text data into meaningful functional units.
- Build a good embeddings model which can encapsulate sub-word information.
- Use it to calculate inter-manuscript distances.
- Plot the Trees based on the distance matrix constructed.

“The Difficulty Lies Not So Much In Developing New Ideas As In Escaping From Old Ones.”

Roadmap

- Introduction
- Dataset
- Experiment Setup
- Methodologies
- Results and Discussion
- Conclusion and Future Work

Introduction

Motivation

Tracing the root of a text *i.e.,* the original version of the text from a given set of manuscript specimens, by inferring phylogenetic trees has been a topic of interest in philological studies.

Existing methods face meaning conflation deficiency due to the usage of lexical similarity based measures.

Distance matrix construction is inherently flawed!

We need a method which can increase the distance between minutely dissimilar manuscripts.

Phylogenetics

Phylogenetics is defined as the task of creating a tree which represents a hypothesis about the evolutionary ancestry of a set of genes, species or any other taxa.

It is the study of evolutionary history and relationships among various taxa.

A Taxon represents a group of one or more manuscripts written in Sanskrit in our case, where we analyze how the manuscripts are related to each other.

Phylogeny - a diagrammatic hypothesis about the history of the evolutionary relationships of a group of manuscripts of the same text.

Why Computational, though?

The computational purview of our problem deals with developing new methodologies for the estimation of the said trees or coming up with methods which can improve the tree construction using the currently available methods.

Hard Reality: Nearly impossible to trace the actual tree with 100% certainty.

Basic Assumption: Not all the traits can be considered while creating a phylogeny. So, we generate a tree based on ‘observable’ traits.

Computers can observe traits faster than we do. Although not always as accurately as we do, but definitely faster.

Word Embeddings

Existing methods do not take into account the sense of a word of even a loose distributional similarity of words.

An increasing boom on large-scale pre-trained word embedding models - fastText, BERT, ELMo, GPT etc. have attracted considerable attention in the field of NLP.

Word embeddings have demonstrated their effectiveness in storing valuable syntactic and semantic information.

A wide range of applications have reported improvements upon integrating word embeddings, including machine translation, syntactic parsing, text classification and question answering, to name a few.

Key Question

Can **word embeddings** help build more a**ccurate phylogenetic trees** from multiple versions of a text in the form of manuscripts?

Dataset

Kāśikāvrtti (KV) Dataset

For distance matrix generation, we focus on specific portions of the KV. We collect seventy different manuscript versions of the KV on Aṣṭādhyāyī (AST) 2.2.6.

We perform cleaning and manual analysis with the help of philologists.

These versions were available in different parts of the country from where we accumulated them in a single repository.

We perform our experiments only on the text of the KV on the AST 2.2.6.

Raw Corpus for Embeddings

We download the Sanskrit Wikimedia and collate all articles in a single corpus.

Add Glosses and Example sentences from the Sanskrit Wordnet to this corpus.

We use various online resources for Sanskrit Text and append to this corpus.

We perform cleaning for this corpus by removing any other ASCII characters apart from the Devanagari script.

The final cleaned corpus used for creating embeddings contains 5,38,323 lines.

Experiment Setup

Preparation

The **Neighbor Joining method and the UPGMA** method are both distance-based methods.

They **require a distance matrix** which specifies the distance between the Taxa being used to populate the phylogeny.

For our experiments, we divide the KV data into different functional units. The functional unit division in KV depends on the type of sutra.

Functional Units

The sutra that we use for our experiments, namely AST 2.2.6, is of the type ‘vidhi’ *i.e.,* this type of sutra prescribes either a verbal element or an operation.

The functional unit division of the text of the KV on this type of sutra is as follows:

1. The sentence explaining the meaning of the words in the sutra.
2. Examples

We compare each functional unit only with its counterpart from the various manuscript versions.

Word Embeddings based Models

We choose FastText for training the word embeddings and obtaining vectors as it utilizes subword-level information within the text.

Sanskrit is morphologically rich and derivationally highly productive language.

To capture the morphology and semantics within each word, we also need to take into account the sub-word level information.

We train the models with the following hyperparameters.

We create these models based on 100 and 50 dimensions due to a limited amount of the corpus collected.

Methodology

Approaches

We use two approaches for constructing the inter-manuscript distances.

- We calculate each inter-manuscript distance by averaging over ‘Unit Distances’ based on:
  - The **baseline approach** utilizes various lexical similarity based measures and later, we also provide weights to them, using empirical approaches, to increase their efficiency.
  - In our approach, we use **word embedding based** models and compute **distances** using vectors obtained from them.
    - Cosine Distance.
    - Angular Cosine Distance (angular cosine distance distinguishes nearly parallel vectors better).

Baseline Approach

We compute the distance between a text by averaging over each ‘Unit Distance’ present in a text (which in our case is a “*sutra+KV”*).

We generate three inter-manuscript distance matrices based on the methods described below:

Weighted Lexical Distance = (NED + CoD + JWD) / 2

NED - Normalized Edit Distance

CoD - Cosine Distance *i.e.,* 1 - Cosine Similarity

JWD - Jaro- Winkler Distance

Word Embeddings based distance

We experiment with two different approaches under the umbrella of word embeddings.

We compute the *cosine distances* between pairs of all words which belong to the same functional unit of manuscript labels being compared.

We average over all the word pair scores and find the functional unit distance.

We average of all the functional unit distances to find the inter-manuscript distance which are then used to construct the distance matrices.

We perform the same steps and obtain distances using *angular cosine distances*.

Tree Construction Methodologies

Neighbour-Joining is a bottom-up clustering method for the creation of phylogenetic trees. It applies general data clustering techniques to sequence analysis and uses genetic distance as a clustering metric. The simple version of the neighbour-joining method produces unrooted trees, but it does not assume a constant rate of evolution across lineages.

The Unweighted Pair Group Method with Arithmetic mean **(UPGMA) method** produces rooted trees and requires a constant-rate assumption, i.e. they assume an ultrametric tree in which the distances from the root to every branch tip are equal.

Results

Observations

We constructed 18 different trees by combining various baseline distance measures, and our approaches by combining them with two different tree construction methodologies.

Resultant Trees:

Best of Embeddings measure and Best of Baseline measure

We observed that in all the cases word embeddings based approaches produce trees which were closer to the expected output of this tree via the help of philologists. We also observed that manuscripts which did not contain any text in their respective functional units were grouped closer to each other *i.e.,* belonging to the same clade.

Types of Variations Involved

Omission (Om.): absence of a word.

Addition (Add.): presence of an additional word

Change of word (CW): lexical changes in the word, generally due to the opinion of the scribe who created the manuscript variant.

Change in the place of a word (CPW): change in the positioning of a word among the functional unit in a text.

Examples

Tri39 - नञ् - नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। - न ब्राह्मणो अब्राह्मणः। अवृषलः।।;

Omission (Om.):

Tri37 - नञ् - नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च भवति । - न ब्राह्मणो अब्राह्मणः। अवृषलः।।;

Addition (Add.):

Bhu1 - नञ् - नञ् इत्येतत् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। - -- अब्राह्मणः। अवृषलः।।;

Some more!

Tri39 - नञ् - नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। - न ब्राह्मणो अब्राह्मणः। अवृषलः।।;

Change of word (CW):

Th1 - नञ् - नञ् सुबन्तेनसमर्थेन सह समस्यते तत्पुरुषश्च समासो भवति। - न ब्राह्मणो -। अवृषलः।।;

Change in the place of a word (CPW):

Hp1 - नञ् - नञ् सुबन्तं समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। - न ब्राह्मणो अब्राह्मणः। अवृषलः।।;

Conclusion and Future Work

Summary

We presented a novel word embeddings based approach to create inter-manuscript distances and hypothesize functional units as a part of the text.

We devised a baseline approach for drawing a comparison.

We present our approach which utilizes word embeddings captured from a generic embeddings models of Sanskrit corpus collected from various sources.

We construct trees using two different methods based on distance matrices obtained via nine different approaches.

We observe that angular cosine distance provides a better distancing mechanism given small changes in the text based on vectors values from the embeddings.

Future Work

In future, we would to like to experiment with contextual embeddings to build the similar trees and check for accuracy.

We would also like to accumulate more raw Sanskrit corpus to build a larger models which can provide us better distributional similarities.

#notetoself: beg for more text corpus! :))

We would also like to experiment with different methods of tree construction which may be able to utilize word embeddings based approaches.

Acknowledgements

Prof. Eivind Kahrs

Dr. Irawati Kulkarni

Dr. Nilesh Joshi

University of Cambridge, United Kingdom.

IIT Bombay, India.

British Academy, United Kingdom.

Rashtriya Sanskrit Sansthan, New Delhi, India.

All the manuscript libraries.

Thanks

Questions?