Deep Learning based Unsupervised POS Tagging for Sanskrit

Deepanshu Aggarwal *2015IPG-113*
Kushal Chauhan *2015IPG-045*
Prakhar Srivastava *2015IPG-063*

Group No.: 26
Mentors: Prof. Anupam Shukla, Dr. Joydeep Dhar

Background:

Processing of natural languages in computer science generally requires parts-of-speech information of the language corpus under study. This parts-of-speech data provides information about the context and usage of the respective words and has been proven to be incredibly useful in Natural Language Processing applications. For languages like English, French and Spanish, this POS information is inferred from a given untagged text to facilitate further NLP tasks. The models used for POS tagging are trained in a supervised manner from large human tagged corpora, and thus achieve excellent performance. The unavailability of tagged corpora for languages like Sanskrit, act as a bottleneck for these traditional supervised learning algorithms as they typically require rich datasets. To overcome these problems, rule based techniques have been employed for POS tagging in Sanskrit, but these achieve moderate performance at best and are very difficult to create as they require extensive knowledge about the grammar and semantics of the language.

Motivation:

Sanskrit is an ancient Indian language whose annotated POS corpora are not readily available to go with supervised learning methods. Though much work has been done for other languages like English, Chinese and French, not much attention has been given to the most scientific and logical language, i.e. Sanskrit. The work that has been done in the domain of Sanskrit POS tagging do not employ novel deep learning methods which have shown much promise recently.

There are many applications of POS tagging which further contributes in the motivation. Some of the applications are given below:

Language Modelling: Language modelling is the process of restrictive appropriation on the recognition of the i^{th} word in an arrangement, given the personalities of every single past word.

Machine Translation: Machine translation is the interpretation of content by a PC, with no human contribution.

Word Morphology: Morphology, in phonetics, is the investigation of the types of words, and the courses in which words are identified with different expressions of a similar dialect.

Literature Review:

The Singular Valued Decomposition approach along with clustering has been used by Schütze (1995) for unsupervised POS tagging. In [6], this approach was modified by employing 'two step SVD' or SVD2 approach, for extracting the latent features from the corpora, followed by K-Means clustering. This implementation achieved more tagging accuracy than the previous method with lesser computational cost.

In the paper by Hammad Ali [9], unsupervised POS tagging was carried out for Bangla language by using a Hidden Markov Model. The states in the HMM were the POS tags and the observed symbols were the words in the corpus. A sequence model was thus constructed for Bangla text, by finding the HMM parameters using the Baum-Welch algorithm.

In a paper by Ammar et al [10], a Conditional Random Field Autoencoder was used to predict latent representation of the input. The latent representation thus generated is converted back into the input by a generative model. A simple categorical distribution over the input vocabulary was used as the input to the CRF Autoencoder to generate its latent representation. This approach enables construction of feature rich representations without resorting to independence assumptions, such as those used in Hidden Markov Models. Competitive results were obtained by employing this approach on problems of POS inductions and bi-text word alignment.

In a proposed approach, instead of using a simple multinomial distribution over the vocabulary for generating word vectors, a multivariate gaussian distribution is used to produce word embeddings for each word token. Specifically, two types of word embeddings are used to capture syntactic similarities:

1. Skip Gram embeddings - These embeddings predict surrounding context words given a current word.
2. Structured skip gram embeddings - These embeddings are an extension of Skip Gram embeddings which also take into account the relative position of words for a given context while generating word vectors.

This approach was tested for unsupervised POS induction using Hidden Markov Models and Conditional Random Field Autoencoders and achieved competitive results [11].

A rule-based POS tagger has been created for Sanskrit in [8]. All the inferred rules are stored in a database. Sanskrit sentences are processed with a specific goal to be allocated a POS tag. For this, the longest addition is sought from postfix table and labels are allocated. 100 wordswith 15 labels are utilized for testing pusposes. 90% accuracy was achiever utilizing this Rule-based approach.

A POS tagger for Bengali has been actualized using deep learning in a paper by Kabir et al [7]. A Deep Belief Network (DBN) was used to construct many layers of latent representations.

A DBN is a deep learning architecture where multiple Restricted Boltzmann Machines (RBMs) are stacked one over the other where the hidden layer of an RBM acts as the visible layer for the next. Some essential features of words in the context of POS tagging were identified such as:

1. Word Length
2. Suffix/Prefix of the word
3. Information about POS tag of the previous words

Research Gaps:

1. Work done in the domain of POS tagging for Sanskrit is very less. Moreover, the tagging that has been done is supervised in nature which requires large well tagged Sanskrit corpora [3][4][8]. As the amount of human tagged corpus for Sanskrit is limited, such approaches are unable to achieve high accuracies.

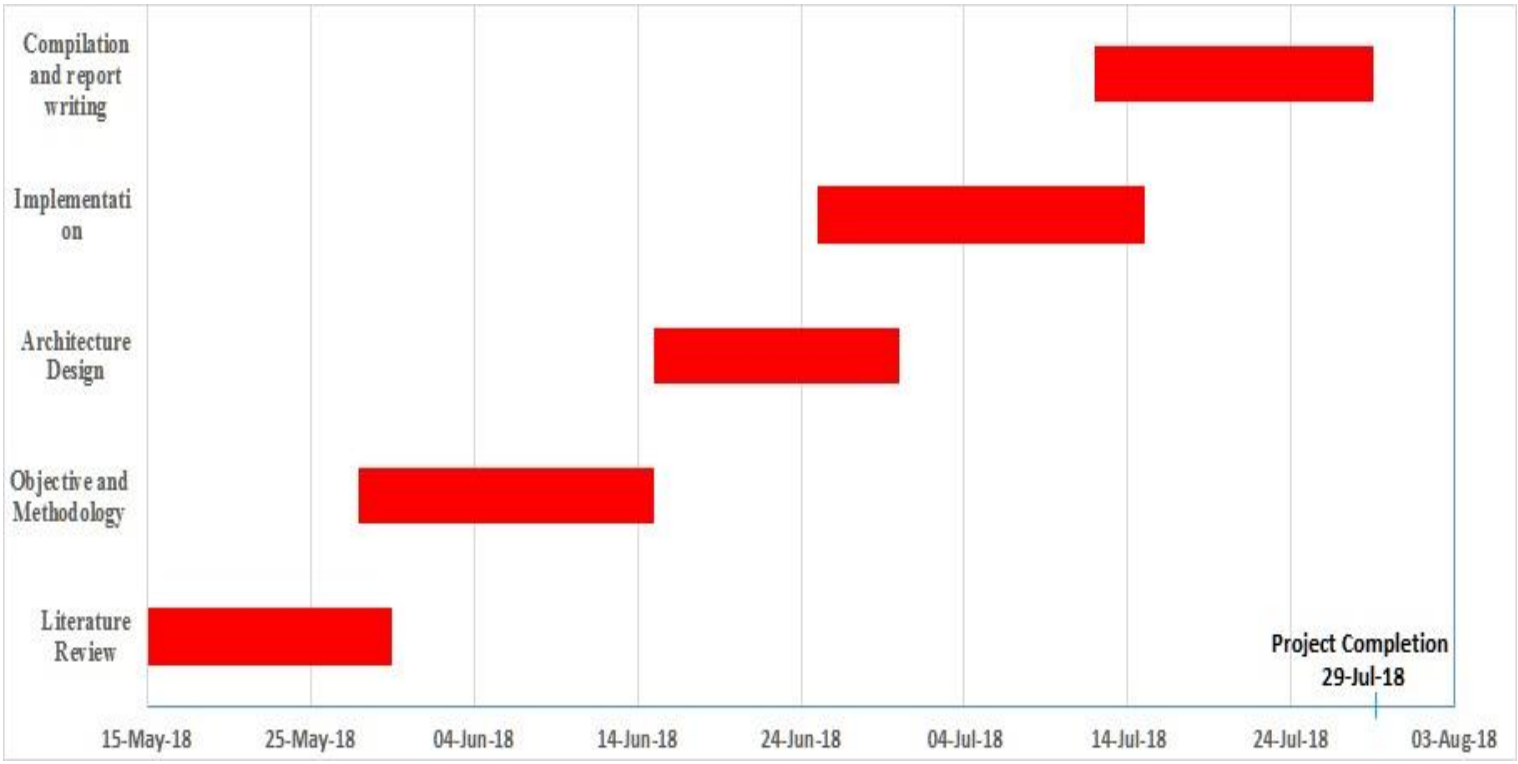
2. Approaches used for POS tagging in Sanskrit have not employed modern deep learning architectures which show much promise [4]. Significant improvements in POS tagging have been observed in other foreign languages and Indic languages such as Bangla [7]. However, the potential of employing novel methods of deep learning in the domain of Sanskrit POS tagging is yet to be explored.

3. The use of Hidden Markov Models in the task of POS tagging has shown some promise [9]. However, the use of HMM requires resorting to independence assumptions which generally are not true in language modelling tasks. Words occurring in natural language are generally dependent on words that occur previously, and influence words that may occur in the future.

4. A deep learning architecture such as Deep Belief Network, which has been used for the task of POS tagging in [7], has to be given inputs which explicitly capture relevant features of previous occurring words, to accurately predict the POS tag of the current word. Feature engineering of this kind is not reliable, but it has to be used in models which lack memory of past inputs.

Problem Statement:

The objective of the project is to apply novel unsupervised deep learning methods to perform POS tagging for Sanskrit to improve upon the current state of the art.



Bibliography:

[1] M. Gopal, D. Mishra, and D. P. Singh, Evaluating tagsets for sanskrit, in Sanskrit Computational Linguistics, pp. 150–161, Springer, 2010.

[2] K. Dinesh and J. G. Singh, Part of speech tagger for morphologically rich indian languages: A survey, International Journal of Computer Application, vol. 6, no. 5, pp. 32-41, 2010.

[3] R. M. Prashanthi, M. S. Kumar, and R. R. Sree, Pos tagger for sanskrit, International Journal of Engineering Sciences Research-IJESR, vol. 4, 2013.

[4] S. Adinarayanan and N. S. Ranjaniee, Part-of speech tagger for sanskrit: A state of art survey, International Journal of Applied Engineering Research, vol. 10, no. 9, pp. 24173–24178, 2015.

[5] Fahim Muhammad Hasan, Naushad UzZaman and Mumit Khan, Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla, In Advances and Innovations in Systems, Computer Sciences and Software Engineering, pp.121-126, Springer 2007

[6] Michael Lamar, Yariv Maron, Mark Johnson and Elie Bienenstock. SVD and Clustering for Unsupervised POS Tagging, Proceedings of the ACL 2010 Conference Short Papers, pp. 215–219, 2010.

[7] Md. Fasihul Kabir, Khandaker Abdullah-Al-Mamun and Mohammad Nurul Huda, Deep Learning Based Parts of Speech Tagger for Bengali. International Conference on Informatics, Electronics and Vision, 2016

[8] Namrata Tapaswi and Suresh Jain. Treebank based deep grammar acquisition and part-of-speech tagging for Sanskrit sentences. In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on, pp. 1-4, IEEEE, 2012

[9] Hammad Ali. An Unsupervised Parts-of-Speech Tagger for the Bangla language, Department of Computer Science, University of British Columbia. 2010.

[10] Waleed Ammar, Chris Dyer and Noah A. Smith. Conditional Random Field Autoencoders for Unsupervised Structured Prediction, NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing, Systems, Volume 2, pp. 3311-3319, 2014.

[11] Chu-Cheng Lin, Waleed Ammar, Chris Dyer and Lori Levin. Unsupervised POS Induction with Word Embeddings, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp. 1311–1316, 2015.