Demo: Part-of-Speech Tagger for Bhojpuri

Srishti Singh and Girish Nath Jha
Jawaharlal Nehru University,
New Delhi, India
{singhsriss, girishjha} @gmail.com

Abstract

This paper is a demonstration of a POS (Part-of-Speech) annotation tool created for Bhojpuri, a lesser resourced language. Bhojpuri is a popular Indian language and spoken by more than 33 million speakers (census 2001) in India. The digital platform the availability of a good POS tagger is an important requirement for language resource creation and the POS tagger discussed here is one of the initial experiments aiming at language resource creation for Bhojpuri. The tagger was created as part of dissertation work and is based on the BIS (Bureau of Indian Standards) annotation scheme. Tagger performs decently on other varieties of Bhojpuri as well because of the variety of corpus data collected from different sources. The average accuracy achieved by the tool, so far, is 88.6% for general domain.

Key words: Annotation Tool, Bhojpuri POS tagger, Demonstration

1. Introduction

Bhojpuri is language of 33 million speakers majorly in U.P. and Bihar state of India and other countries like Nepal, Bhutan Mauritius, Fiji, Guyana etc. Although, Bhojpuri has gained a lot of attention through Bhojpuri cinema worldwide, it is still struggling for its recognition as a standard language and has no technological resource. Therefore, the motivation behind creating 'Bhojpuri POS tagger' is to bring it to the Digital platform and anticipating other language resource for the language in future. The present POS tagger is one of the pioneering works in this field which is calculated to have an average accuracy of 88.6% which can be found on the following website: (http://sanskrit.jnu.ac.in/bhopos/index.jsp)

2. Bhojpuri POS Tagger

2.1 Tagger description

The general domain representative Bhojpuri Corpus with approx. 192k tokens was created as part a Research work. This is the first big corpus for Bhojpuri. The corpus data is collected from some manually transcribed Bhojpuri folk children stories, websites for literary article, news, magazines, literature etc like bhojpurika.com and anjoria.com with majorly literature, entertainment, politics, sports and blogs etc. The data for corpus creation is collected both manually and semi-automatically using ILCrawler and Sanitizer for collection and corpus cleaning. (Singh, 2015b).

A two-tier hierarchical tagset for Bhojpuri was designed in this endeavour modelled on BIS standards[1] (annotation scheme for all Indian languages). The tagset initially had 33 tags as reported in Singh (2014) but it latter included a new tag label called echo-before (Ech_B) for phrases like 'adalA-badalI[2]' found in both Hindi and Bhojpuri where the second word means *to change* whereas the first word it the echo of the second preceding it (Singh, 2015). The tagger is trained with Support Vector Classificatory model (SVM) for its excelling performance of big data (Giménez, 2004).

2.2 Tagger Architecture

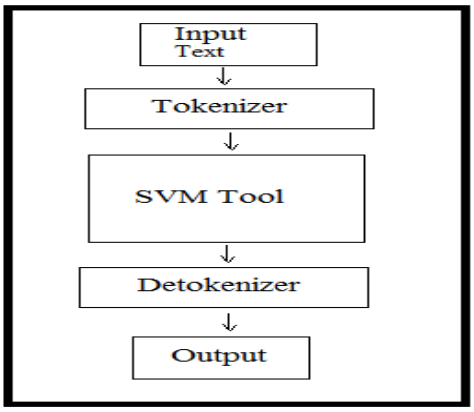

Fig. 1 Architecture of the tagger

Validated raw Bhojpuri corpus is the input for the tool which is first pre-processed and tokenised. The tokenised corpus serves as the input for the SVM machine and the POS tagging is done. The Tagger output s also in tokenised form, therefore, a de-tokenised is used for post-editing is used before displaying the tagged output. The training and test data used for testing is in 80-20 ratio as per the annotation standards.

2.3 Tagger output

Initially the training was performed on a set of 30k tokens and the accuracy of the tagger was calculated to be ranging between 74-85% for random set of data. The latest report on tagger

[1] BIS Guideline: (http://sanskrit.jnu.ac.in/ilciann/index.jsp)

[2] **Itrans** is used for Romanisation the Bhojpuri text throughout the paper.

shows the accuracy of 88.6% when trained on 90k token (Singh, 2015). Currently, the tagger is under development and the training size in being increased along with the size of the Bhojpuri corpus. The Hindi POS tagger trained under ILCI[3] project exhibits an accuracy of approx. 94% at present (Ojha, 2015) which can be found at (sanskrit.jnu.ac.in/pos/index.jsp).

3. Tagger evaluation

Despite belonging to the same language family and sharing much common linguistic features, the use of classifiers, ergative markers, imbedded demonstratives and lexical ambiguity are found in Bhojpuri which are not present in Hindi language. The ambiguity noticed in the corpus is one major challenge for the machine learning. There were ambiguous tokens found in the corpus claiming up to four possible tags for single token as well as four realizations for one tag in different contexts. At the level of POS category, the tagger encountered maximum issues with auxiliary in serial verb constructions; noun & adjectives in conjunct verbs, and ambiguous tokens.

One example of **homophones** cited from Singh (2015b) where 'ka' and 'ke' tokens were often confused with their part of speech category in different contexts. From the corpus it was found to belonging to three possible categories namely subordinator, postposition and auxiliary verb. For example:

1. (kAhe **ke)** sabale manjUra rahale (BHO)
   because all agreed to it (Eng)

2. (lA **ke)** de dA (BHO)
   bring it for him (Eng)

3. hama sUraja DUbe **(ke** bAde) jAiba (BHO)
   I will go only after the sunset (Eng)

The token 'ke' is used as part of *kAhe ke as* subordinator in example 1, *lA ke* as an auxiliary verb in example 2 and *ke bAde* as part of complex postposition in example 3.

Similarly, example of varied **realizations of single token** 'aura' (and) is considered. The conjunction *aura* is represented as '*aura', 'A', 'a',* and '*au'* throughout the corpus as reported in Singh (2014). Moreover, other tool related challenges.

4. Development and Future work

The technological advancement is important for the expansion of a language and resource creation helps retaining and updating the orally transferred knowledge and literature, with time. The present Bhojpuri tagger is an initiative for providing a platform to Bhojpuri and for other NLP tools to come into existence.

The present tagger is under development and both the tagger accuracy and corpus size is being worked upon so that other higher level technological resources can be developed based on the efficiency of the tagger, adding on to the advancement of Bhojpuri.

5. Acknowledgements

We are thankful to ILCI consortium project for providing technical help for the experiment and LREC for considering the paper for demo presentation.

Reference

Choudhary, N. and Jha, G. N. (2011). *Creating Multilingual Parallel Corpora in Indian Languages*. In Proceedings of Fifth Language Technology Conference, Poznan, Poland.

Giménez, J. and Màrquez, L. (2004). *SVMTool: A general POS tagger generator based on Support Vector Machines*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal.

Ojha, A., Behera, P. And Singh, S. (2015). *Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri.* In Proceedings of seventh Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland.

Singh, S and Banerjee, E. (2014). Annotating Bhojpuri Corpus using BIS scheme. In Proceedings of the second Workshop on Indian Languaage Data: Resource and Evaluation (WILDRE). In Proceedings of Ninth International Conference on Language Resource and Evaluation (LREC'14), Reykjavik, Iceland.

Singh, S. (2015a). *Challenges in Automatic POS Tagging of Indian Languages- A Comparative Study of Hindi and Bhojpuri*. Unpublished M.Phil Dissertation submitted to Centre for Linguistics, Jawaharlal Nehru University, New Delhi.

Singh, S. (2015b). *Statistical Tagger for Bhojpuri: Employing Support Vector Machine*. In Proceedings of Forth International Conference on Advances in Computing, Communications and Informatics (ICACCI'15), Kerela, India.

[3] ILCI- **Indian Languages Corpora Initiative** Consortium Project headed by Jawaharlal Nehru University (Choudhary and Jha, 2011)