Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri

Atul Kr. Ojha Pitambar Behera, Srishti Singh and Girish Nath Jha
Jawaharlal Nehru University, New Delhi
(shashwatup9k, pitambarbehera2, singhsriss & girishjha)@gmail.com

Abstract

The present paper discusses the training and evaluation of the CRF and SVM algorithms for Indo-Aryan languages: Hindi, Odia and Bhojpuri. For annotation of the corpus, we have used Bureau of Indian Standards (BIS) annotation scheme which is a common standard of annotation for Indian languages. The main objective of the paper is to provide an idea of the error pattern and suggestions following the same algorithms. The experiment is conducted with 90k tokens training and 2k tokens test data each, for ease of comparison among languages. In the evaluation report, we focus on each tool (SVM and CRF++) at the level of accuracy, error analysis of the tools, the error pattern and common error of the system. The accuracy of the SVM taggers ranges between 88 to 93.7 % whereas CRF ranges between 82 to 86.7%. CRF performs less qualitatively than SVM for Odia and Hindi which is not true for Bhojpuri. In this study, we have observed that languages having more variations are suitable for CRF in comparison to SVM.

Keywords: SVM tool, CRF++, Hindi, Bhojpuri, Odia, statistical POS tagger, evaluation.

1. Introduction

The paper presents a comparative evaluation of reports obtained as result of the statistical POS tagging tools based on Support Vector Machine (SVM) and Conditional random Fields (CRF) algorithms for Indo-Aryan (IA) languages. The languages in this endeavour are Hindi, Odia and Bhojpuri. The paper provides an overview of the study conducted in Hindi and Odia. It also provides the scope of development for Bhojpuri; which is relatively new in this area of technological advancement. The paper demonstrates the issues and challenges encountered during the course of training and evaluation. The error pattern for each language has also been studied in its linguistic aspect.

1.1. Areal Features

Hindi, Odia and Bhojpuri belong to Indo-Aryan language family with SOV word order. Hindi is one of the official languages in the Republic of India. Odia is one of the classical languages, spoken in the eastern region of India. Apart from inheriting most of the linguistic features from the IA group, it also has some Dravidian features (Patnaik, undated); due to its geographical location. It is also spoken in the neighbouring states[1] of Odisha (formerly Orissa), some parts of West Bengal, Chattisgarh, Jharkhand, Andhra Pradesh and by the overseas population in the U.S. and U.K. and in some other countries.

Bhojpuri is the language of Northern India spoken in the east of Uttar Pradesh and Bihar along with some other foreign countries like Mauritius, Nepal, Surinam etc.

1.2. POS Annotation

Parts of Speech tagging is a process of assigning grammatical categories to each word in a running text. Being a morpho-syntactic process the context-based meaning of the word is considered during manual annotation. We have used SVM and CRF tools for automatic taggers.

Support Vector Machine (SVM)

SVM is a classification and regression algorithm which is based on the statistical learning theory developed by Vapnik and his team, in 1995. Support Vector-based classifiers are capable of potentially handling more than two classes of variables on both linear and non-linear planes.

$$\mathbf{w}^* = \sum_{i=1}^{N} y_i \alpha_i^* \mathbf{x_i}$$

Where α is a given vector and w is the feature component for maximum margin hyperplane.

Conditional Random Fields (CRF)

CRF is a statistical tagging model developed by Charles Sutton. It is a probabilistic model working as follows:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{\Psi_A \in G} \exp\left\{ \sum_{k=1}^{K(A)} \lambda_{Ak} f_{Ak}(\mathbf{y}_A, \mathbf{x}_A) \right\}$$

Where G is taken as a factor graph over y, then p (y|x) is a Conditional random field if the distribution factorizes according to G for any fixed x (Agarwal and Mani, 2011).

The data is trained upon SVM Tool (version 1.3.2) for SVM[2] and CRF ++ (version 0.58) on C language for CRF[3].

1.3. Typological Features

In this sub-section, the common typological features of the IA languages with respect to classifiers, inflection, agglutination, PNG (person, number and gender), particles and verbal constructions are discussed. The cross-linguistic comparison has been discussed as some of the features create issues in POS annotation (see section 5).

1.3.1. Classifiers

The classifiers are very prominent in Odia. Like its

[1] www.ethnologue.com

[2] http://www.cs.upc.edu/~nlp/SVMTool/

[3] https://taku910.github.io/crfpp/

sister languages namely, Bangla and Assamese, Odia also has nominal and numeral classifiers but Bhojpuri does not use classifier with the nouns. This feature is absent in the Hindi.

For example: numeral classifiers

Hindi:	eka	do	tIna
Odia:	eka-**Ti**	dui-**TA**	tIna-**Ti**
Bhojpuri:	eka-**go**	du-**go**	tIna-**go**
Free translation:	one	two	three

1.3.2. Inflection and Agglutination

Hindi and Bhojpuri are inflecting languages in which a new word is formed with affixes inflected with it whereas Odia is morphologically an agglutinating language (Mohapatra, 2010 and Jena et al, 2011). Case markers and suffixes (numeral, singular and plural) in nouns agglutinate with the nouns in Odia (Behera, 2015).

For example:

loka-ra 'man's' (genitive case marker)
loka-ku 'to the man' (dative case marker)

1.3.3. Person/Number/Gender

There are more verbal agreement restrictions in Hindi than in Bhojpuri and even less in Odia. Hindi follows a very strict gender agreement with the verb and has a clear demarcation of the masculine and feminine genders. In Bhojpuri, the morphological gender is more prominent than the grammatical gender.

For example:

Hindi:	maiM	Ay-A (M)/Ay-I(F)
Bhojpuri:	ham	a-il-I (M/F)
Gloss:	I	come (M/F)

Free translation: I came.

In Odia, verbs agree with the person and number of the nouns and can appear with a covert subject. The concept of grammatical gender is not present, but the morphological gender is lexically marked in adjectives for the gender of the nouns they qualify.

For example:

(a) Odia: muM kAli Asili
Gloss: I-1SG Yesterday come.PAST.PFV
Free translation: I came yesterday.
(b) Odia: sundarI jhia
Gloss: Beautiful girl (Lexical gender)
Free translation: Beautiful girl.

1.3.4. Particles

The arrangement of emphatic particles in these languages is different. They are relatively free-floating in all. Hindi and Odia emphatic particles always take a morphological form. Bhojpuri particles are capable of occurring both independently as a separate word unit and as an inflection to the parent category.

For example:

Hindi: mai abhI bAzAra se hI AyA huM
Gloss: I now market from **PRT** come be
Odia: muM bartamAna hiM bajAraru Asili
Gloss: I now **PRT** market come-PST-PFV
Bhojpuri: ham abahiyeM bajAre se ailI ha
Gloss: I now-**PRT** market from come be

Free translation: I have just come from the market.

1.3.5. Verbal Constructions

There are some other typological differences in these languages which are very significant to this work because in due course of POS tagging the morpho-syntactic nature of languages are considered and the verbal constructions in Odia are very different from that of Hindi and Bhojpuri.

Following BIS guideline, the tagset devised for Hindi and Bhojpuri has only two verb categories namely main and auxiliary verbs but for Odia unlike the other two, there are five tag-level categories (main, auxiliary, finite, non-finite, infinitive and gerund).

In Odia, the Tense Aspect and Mood (TAM) and person/number (PN) morphemes appear in agglutinated form with the verbs. For example, *-bA* represents infinitive form of verb while *–kari* marks non-finite forms. Similarly, *-ib*, and *–il* refer to future and past respectively. This leads to the categorization of the main verb into five as proposed in the BIS scheme.

2. Literature Survey

In this section, a brief literature survey of POS annotation of these languages has been presented.

2.1. Hindi

Several statistical models like HMM, ME, CRF were tested in 2008 and morphology driven (2006) Parts of Speech Taggers have been trained for Hindi. The morphology driven tagger proposed by Smriti Singh, worked on decision tree algorithm with an accuracy of 93.45%. The Maximum Entropy based Hindi tagger by Aniket Dalal reached the overall accuracy of 88.4% over 10 runs. This also showed that the corpora size affects the training results. The Conditional Random Fields based tagger by Agarwal and Mani in 2006 made use of morph analyser for training. It was trained on 21k words with the best feature set having 82.67% accuracy, using CRF++ tool. An improved HMM tagger was developed using stemmer for suffix matching and a pre-processor with an accuracy rate of 93.12%.

Another tagger was proposed by (Ray et al., 2003) using constraint propagation which was based on ontological information, morphological and lexical rules, and capable of capturing four local dependencies of Hindi but was not tested.

2.2. Odia

Das and Pattnaik (2014) have proposed a Single Neural Network-based parts of speech tagger for Odia language. The tagger has been selected empirically with the fixed length of context, initially, which was later corrected by forward propagation and transferred using 'feed-forward technique' for multiple layers. This has been reported to have reached accuracy of 81 %. Then, Das et al., (2015) developed an SVM Tagger with reported accuracy of 82% with 10k tokens training data. The tagset used consisted of only five tags, along with careful handling of prefixes and suffixes.

2.3. Bhojpuri

Initiations are being taken for building resources for Bhojpuri. The POS tagset for Bhojpuri has recently been developed (Singh and Banerjee, 2014) following BIS guidelines. The POS tagger for the language has been developed by a group of Computational linguists at JNU and it was trained on the data of 90k tokens with accuracy ranging between 82 to 89% depending upon the domain and type of test data[4].

3. Experimental Set-up

This section has been divided into five sub-categories. It presents the corpus collection, annotation, training methodologies and architecture of the POS taggers.

3.1. Data Collection and Corpus Building

The data for the present experiment has been extracted from the Indian Languages Corpora Initiative (ILCI) project[5] for corpus creation (Jha, 2010). The comparison of the tagger's efficiency for different languages has been made keeping the size of the training set constant. The corpus for Hindi and Odia has been selected from health and tourism domains. On the other hand, Bhojpuri corpus is a web-crawled general domain corpus (Singh, 2015).

3.2. Size of the Corpus

In this experiment, the corpus size of the training and test data for each language is 90k and 2k tokens respectively. The current phase of development reports the total size of the Hindi training data 1,320k tokens, in Odia 236k tokens and for Bhojpuri it is 120k tokens, approximately.

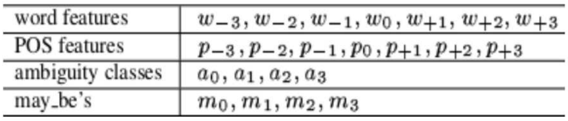
3.3. POS Annotation and Annotation Scheme

Parts of Speech tagging is one of the basic steps for resource creation. The POS tagging of the corpus for all the three languages were done semi-automatically using the ILCIANN[6] POS annotation tool (Kumar et al., 2011) developed under the banner of ILCI project. BIS[7] served as the model for the tagset designed in 2010 as part of the DeitY-sponsored ILCI project. This tagset enabled POS annotation of all the scheduled languages of India using one common scheme.

The tagged data has been manually validated in 2 to 3 folds before the training process started.

3.4. Feature Selection of POS Taggers in SVM and CRF

On one hand, SVM has three modules, namely, SVMTlearn, SVMTagger and SVMTeval. On the other hand, CRF++ has two modules: crf_learn and crf_test. For evaluation of CRF tool, we have adopted the format of conlleval evaluation file scripted in Perl[8] by Sang (2000). The feature selections under discussion are identical. The features for SVM have been selected taking into consideration the word, POS, ambiguity and may_be's. "A may_be's states, for a certain word, that certain PoS may be possible, i.e. it belongs to the word ambiguity class" (Giménez & Marquéz, 2006) and the C parameter has been set to zero. It controls overfitting problems in classifying training data. The features for prefixes and suffixes, dictionary look-up have not been considered and Dratio is (0.005). Dratio helps to build a list of files in repairing the dictionary.



word features	$w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$
POS features	$p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}$
ambiguity classes	a_0, a_1, a_2, a_3
may_be's	m_0, m_1, m_2, m_3

Fig. 1: Template for feature selection of SVM

The template file used for CRF feature selection has been based on Unigram and like SVM, the features for prefixes and suffixes, dictionary look-up have not been taken up.

3.5. Architecture of POS Taggers

The figure demonstrated below presents the user interface architecture of the POS taggers developed for the IA languages. Firstly, the users provide the input text to the POS taggers and select the drop-down button for their respective languages by selecting whether to annotate with the SVM or CRF. Secondly, the taggers internally tokenize the input data and process it. Thirdly, they send the input text to the respective algorithms and process the tagged output. Finally, the tagged data is detokenized and the final output is shown on the display (http://sanskrit.jnu.ac.in/pos/index.jsp).

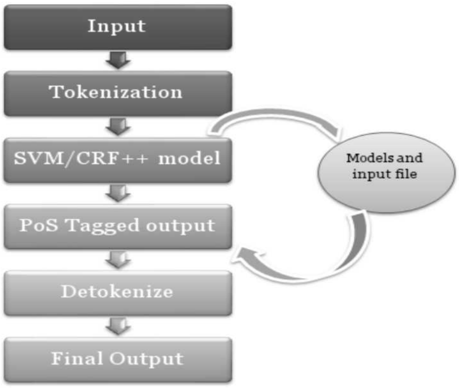


Fig. 2: Architecture of POS tagger

4. Comparative Evaluations of the POS Taggers

This section has been classified into two categories: evaluation of POS taggers and linguistic analysis of errors.

[4] http://rpsonline.com.sg/rps2prod/icacci2015/usb-proceedings/pdf/1570153365.pdf

[5] http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci

[6] http://sanskrit.jnu.ac.in/ilciann/index.jsp

[7] http://www.tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf

[8] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

4.1. Evaluation of POS Taggers

The figure demonstrates the overall tagger accuracy of Hindi, Odia and Bhojpuri which is found to be 82.6, 85.5 and 86.7% following CRF-based taggers and 93.6, 91.3 and 88.5% for SVM-based taggers respectively.

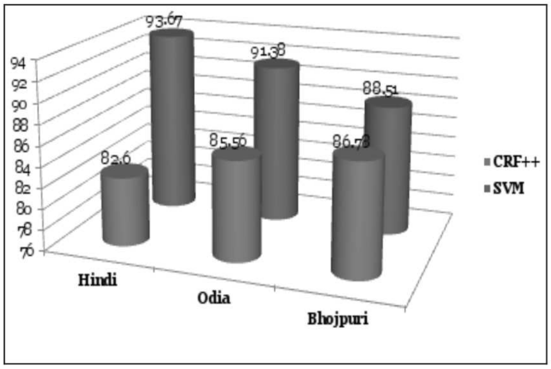


Fig. 3: Overall accuracy (%) of POS tagger

4.2. Linguistic Analysis of Errors

The error report has been established for both SVM and CRF considering the rate of most frequent tagging errors. Fourteen tag labels (N_NNP for proper noun, JJ for adjectives, V_VM for main verb, etc.) out of the whole tagset show the highest rate of error in the report on the basis of precision and recall evaluation metrics.

4.2.1. Rate of Errors for the Taggers on Recall metrics

From figures 4 and 5, the error rate of proper nouns for CRF shows 87.6% in Hindi whereas in Odia and Bhojpuri it shows 54.6 and 22.3 % respectively. Looking into the SVM for proper noun, the error rate of Hindi falls to 18.5% but in Odia and Bhojpuri, error rate is slightly decreased to 46.7 and 12%.

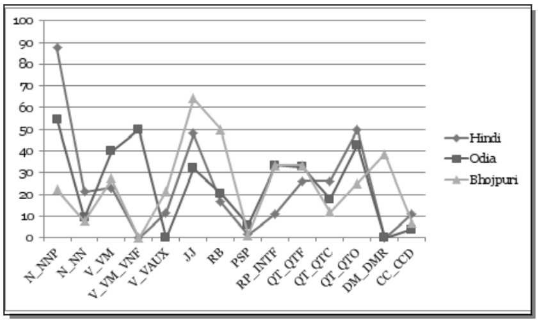


Fig. 4: Error rate (%) of CRF++ on recall

For adjectives, CRF shows much higher error in all the three languages as compared to SVM. The label for infinitival verb (V_VM_VINF) is only present the Odia annotation scheme. In auxiliary verb, Bhojpuri has got around 73.4 % in SVM which falls to 21.4% in CRF and other two (Hindi and Odia) are 23.4 and 0% in SVM and 11.74 and 0% respectively.

So far as Odia is concerned, the minimum error rate in the category of auxiliary verb is owing to the fact that there is no enough number of auxiliary categories present in the testing data. The rate of error for infinitives is 31.5% and 50% for SVM and CRF, respectively.

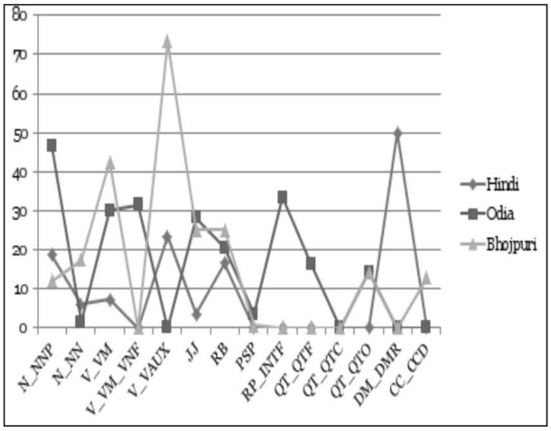


Fig. 5: Error rate (%) of SVM on recall

4.2.2. Rate of Errors for the Taggers on Precision metrics

From figures 6 and 7, the precision report in proper noun shows the higher error rate for CRF-based taggers for all languages except Bhojpuri which is 10.8% in CRF and 22.2% in SVM.

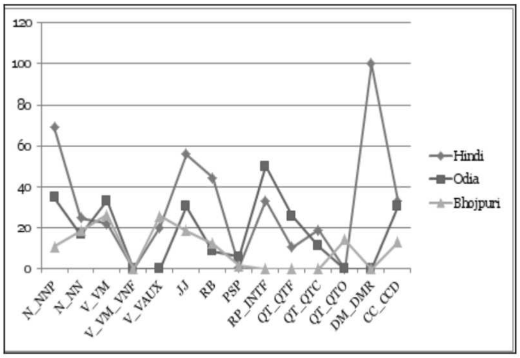


Fig. 6: Error rate (%) of CRF++ on precision

The next high most frequent errors are found in verb category. Main and auxiliary verbs are considered here. The error rate for SVM in Hindi is below 18 % for both the main and auxiliary verbs which is quite high for Odia and Bhojpuri. In comparison with CRF the error rate rises above 20% for all the languages concerned. In adjectives, the rate of error is higher in CRF ranging between 18 to 55 % whereas errors in SVM range between 6.6 to 29.4%.

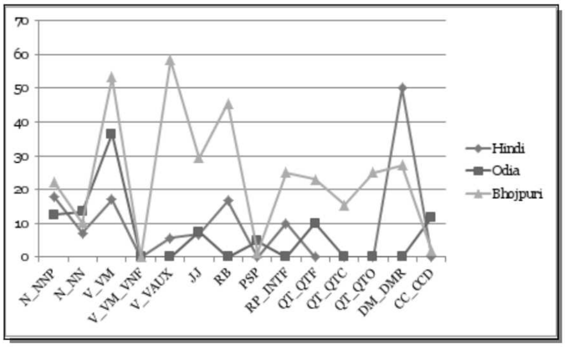


Fig. 7: Error rate (%) of SVM on precision

5. Discussion

The different typological features discussed in the paper involve the following issues in manual and automated annotation. The classifiers and particles can appear both as a morpheme and an inflection to some other morpheme. In this case, the problem arises whether to tag the word as a classifier or to give it the tag of the head category.

For example,

One (English): ek-TA (Odia) and ego (Bhojpuri)

–TA and *–go*, respectively create issues for both the annotator and tagger. Both the words carry a cardinal and a classifier. The POS tag provided will be that of a classifier but not a cardinal. This is also generated by the tool. But in some cases like adverbs in Bhojpuri like *abahiyeM* (now) where, though the word contains a time adverb and a particle, the POS category of the word will be an adverb (temporal according to BIS) but not a particle.

Similarly, the two processes involved are inflection and agglutination. Hindi and Bhojpuri follow purely inflection process which results in the formation of two verb categories (main and auxiliary) but Odia follows both agglutination and inflection in different constructions. They, in the verbal constructions in Odia, lead to the addition of further four subcategories of the main verb.

For instance,

karibA**-ku** 'to do' (infinitive)
kari**-lA** 'did' (finite)
kAma karu**-thibA loka** 'work doing man' (gerundive)
kari**-ki** 'after doing' (non-finite)

In the category of verbal noun, the classifiers also alternate with the verb (*tAra karibA-TA* 'his doing') and postpositions also agglutinate with the verb.as in above infinitive verb. When the numeral classifiers occur in the training data, they are labelled as either classifier or verbal noun because of the ambiguity in deciding the tag. Therefore, the classifiers attached to verbs are automatically tagged as classifiers in some occasions.

Besides, there are other problematic POS categories affecting the performance of the tagger that have already been discussed in the previous section.

6. Conclusion

This experiment initially was conducted with 90k tokens training and 2k tokens test data for ease of comparison among languages. The accuracy of the SVM taggers ranges between 88.51% to 93.67% and CRF between 82 to 86.7% %.

The accuracy has been achieved without applying any external tool (morph analyser, dictionary look-up etc.). The error rate of CRF is comparatively higher than SVM for all the languages except Bhojpuri which is higher in SVM for all tag-level categories. One reason for this feature is due to the high language variation in Bhojpuri which is lesser in Hindi and Odia.

Out of the above discussed categorical labels in sections (4.2.1 & 4.2.2), proper nouns and verbs show higher error rate. For resolving this issue to enhance the performance of the tagger, an NER (Named Entity Recognition) and a morph analyser can be applied. In future, the dictionary look ups, bi-grams and trigrams can also be applied to achieve an enhanced accuracy rate.

Acknowledgement

We are grateful to the DeitY-sponsored ILCI Project for providing corpora and semi-automated annotation tool. Thanks are also due to Jesus Giménez and Eric Tjong Kim Sang for their contributions: SVM and CRF++ tools respectively.

References

Agarwal, H. and Mani, A. (2011). Part of Speech Tagging and Chunking with Conditional Random Fields. *In the proceedings of NLPAI Contest, 2006*. Retrieved from: http://ltrc.iiit.ac.in/nlpai_contest06/proceedings/tilda.pdf. Access date: August 7, 2014.

Behera, P. (2015). *Odia Parts of Speech Tagging Corpus: Suitability of Statistical Models*. M. Phil. Dissertation. New Delhi: Jawaharlal Nehru University.

CRF++: https://taku910.github.io/crfpp/

Das, B. R., & Patnaik, S. (2014). A Novel Approach for Odia Part of Speech Tagging Using Artificial Neural Network. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, pp. 147-154. Springer International Publishing.

Das, B. R., Sahoo, S., Panda, C. S., & Patnaik, S. (2015). Part of Speech Tagging in Odia Using Support Vector Machine. *Procedia Computer Science, Volume 48, 2015, pp. 507-512, ISSN 1877-0509.*

Giménez, J. & Marquéz, L. (2006). SVMTool Technical Manual v1. 3. Retrieved from: http://www.cs.upc.edu/~nlp/SVMTool/SVMTool.v1.3.pdf. Access date: December 12, 2014.

Jena, I., Chaudhury, S., Chaudhry, H., & Sharma, D. M. (2011). Developing Oriya Morphological Analyzer Using Lt-toolbox. *In Information Systems for Indian Languages.* Berlin Heidelberg: Springer. pp. 124-129.

Jha, G. N. (2010). The TDIL program and the Indian Language Corpora Initiative (ILCI). In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta, May 17-23, 2010.

Kumar, R., Kaushik, S., Nainwani, P., Banerjee, E., Hadke, S., & JHA, G. N. (2012). Using the ILCI Annotation Tool for POS Annotation: A Case of Hindi. *In 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012).* New Delhi, India.

Mohapatra, R. and Hembram, L. (2010) Morph-Synthesizer for Oriya Language- A Computational Approach. *Language in India*. Volume 10, September 2010.

Patnaik, B. N. (undated). Oriya as a Typologically Disturbed Language and Some Related Matters. Retrieved from:

http://home.iitk.ac.in/~patnaik/documents/oriya_typo .pdf. Access date: July 20, 2014.

Ray, P. R., V, H., Sarkar, S. and Basu, A. (2003). Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi. *In Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003).* Mysore, India.

Singh, S. (2015). *Challenges in Automatic POS Tagging of Indian Languages- A Comparative Study of Hindi and Bhojpuri.* M. Phil. Dissertation. New Delhi: Jawaharlal Nehru University.

Singh, S. and Banerjee, E. (2014). Annotating Bhojpuri Corpus using BIS Scheme. In *Proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE-2)*, *Ninth International Conference on Language Resources and Evaluation, LREC (2014)*, *Reykjavik, Iceland, May 26-31, 2014*.

Singh, S. and Jha, G. N. (2015). Statistical Tagger for Bhojpuri (employing Support Vector Machine). In *Proceedings of 4th International Conference on Computing, Communication and Informatics (ICACCI, 2015).* Retrieved from: http://rpsonline.com.sg/rps2prod/icacci2015/usb-proceedings/pdf/1570153365.pdf Access date: August 10, 2015.

SVMTool: http://www.cs.upc.edu/~nlp/SVMTool/

Tjong Kim Sang, E. F., & Buchholz, S. (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. *In Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7, Association for Computational Linguistics, pp. 127-13, Lisbon, Portugal, 2000.*