

Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages

Fahim Muhammad Hasan
Center for Research on
Bangla Language Processing
BRAC University
66, Mohakhali C/A, Dhaka
Bangladesh
fahimht@gmail.com

Naushad UzZaman
Center for Research on
Bangla Language Processing
BRAC University
66, Mohakhali C/A, Dhaka
Bangladesh
naushad@bracu.ac.bd

Mumit Khan
Center for Research on
Bangla Language Processing
BRAC University
66, Mohakhali C/A, Dhaka
Bangladesh
mumit@bracu.ac.bd

Abstract

Part-of-Speech (POS) Tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags. POS Tagging is important in various areas of Natural Language Processing. Different methods of automating the process have been developed and employed for English and other Western languages. Some similar work, most of which utilize the stochastic approaches for POS Tagging has also been done in the same area for South Asian languages. We experimented with some of the widely-used approaches for POS Tagging on three South Asian languages, Bangla, Hindi and Telegu, using corpora of different sizes. We observed the performance of the approaches and found the Brill's transformation based tagger's performance to be superior to the other approaches in all of our experiments, though the use of this approach has been very limited until recently.

1. Introduction

Part-of-Speech (POS) Tagging means assigning appropriate grammatical classes (i.e. appropriate Part-of-Speech tags) to each word in a natural language sentence. It has its importance in various areas of Natural Language Processing (NLP) such as Text-to-Speech, information retrieval, parsing, information extraction and linguistic research for corpora [1, 2]. It can also be used as an intermediate step for higher-level NLP tasks such as semantics analysis, translation, and many others [3].

Assigning a POS tag to each word of an un-annotated text by hand is very time consuming. And that is why POS Tagging has become one of the well-studied problems in the field of NLP.

Several different approaches have already been developed and employed for POS Tagging for English and some other western languages.

On the other hand, the amount of work accomplished in the same area for South Asian languages is quite inadequate. Also, most of them have been applying the stochastic methods of POS Tagging.

In this paper, we start by briefly classifying the different POS Tagging approaches. Then we continue by giving a concise overview of the work already done in NLP for English and some South Asian languages. We move on by describing the different models that we use for our experiments. Next, we discuss the corpora that we employ for training and testing the tagging models. We also describe the tagset that we use. Afterwards, we give some examples of the output that our POS Tagging models produce. Then we show how the models perform using the corpora and tagset that we utilize. After that, we analyze the results we find and compare the performance of the tagging models based on different approaches. We conclude with the result that transformation based Brill's tagger is suitable for tagging South Asian languages, Bangla, Hindi and Telegu, using varying corpora sizes up to 25000 annotated tokens. We also propose some future studies that we plan to accomplish.

2. Classification

There are different approaches for POS Tagging. The following figure classifies different POS Tagging models.

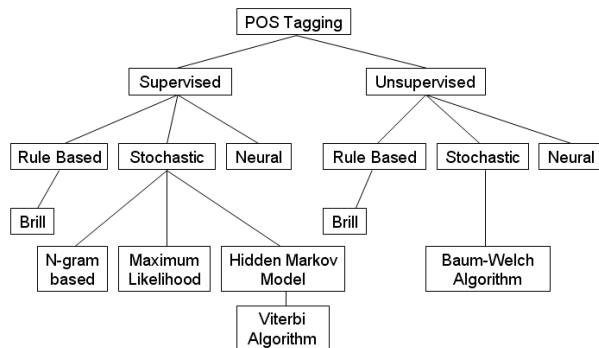


Figure 1. Classification of POS tagging models

Here we give a very brief overview of the different models. A detailed discussion of the models can be found at [3, 4, 5].

2.1. Supervised models

The supervised POS Tagging models require a pre-annotated corpus which is used for training to learn information about the tagset, word-tag frequencies, rule sets, etc. [6]. The performance of the models generally increases with increase in the size of the corpus.

2.2. Unsupervised models

The unsupervised POS Tagging models do not require a pre-annotated corpus. Instead, they use advanced computational techniques like the Baum-Welch algorithm to automatically induce tagsets, transformation rules, etc. Based on this information, they either calculate the probabilistic information needed by the stochastic taggers or induce the contextual rules needed by rule based systems or transformation based systems [6, 7].

Both the supervised and unsupervised models can be further classified into the following categories.

2.3. Rule based and transformation based models

The rule based POS Tagging models apply a set of handwritten rules and use contextual information to assign POS tags to words. These rules are often known as context frame rules. On the other hand, the transformation based approaches use a predefined set

of handcrafted rules as well as automatically-induced rules that are generated during training. Some models also utilize morphological rules [2], capitalization and punctuation, etc. [7].

2.4. Stochastic models

The stochastic models include frequency, probability or statistics. They can be based on different methods such as n-grams, maximum-likelihood estimation (MLE) or Hidden Markov Models (HMM).

HMM-based methods require evaluation of the argmax formula, which is very expensive as all possible tag sequences must be checked, in order to find the sequence that maximizes the probability. So a dynamic programming approach known as the Viterbi Algorithm is used to find the optimal tag sequence [8].

There have also been several studies utilizing unsupervised learning for training a HMM for POS Tagging. The most widely known is the Baum-Welch algorithm [9], which can be used to train a HMM from un-annotated data.

And lastly, both supervised and unsupervised POS Tagging models can be based on neural networks [10].

3. Previous work

In this section we discuss some of the works regarding POS Tagging for some South Asian languages. Information on works done for English and other western languages can be found at [9, 11 - 18].

In [19] the authors report a hybrid tagger for Hindi that uses two phases to assign POS tags to input text, and achieves good performance. In the first phase, an HMM-based tagger is run on the untagged text to perform the tagging. And in the second phase, a set of transformation rules is applied to the initially tagged text to correct errors.

The authors of [20] report a tagging method for Hindi that overcomes the troubles in accurate tagging due to the scarcity of large sized corpora. The technique uses a medium sized corpus of around 15000 words, utilizes a morphological analyzer as well as a decision tree based learning approach to achieve a reported accuracy of 93.45%, which is quite impressive.

For the Telegu language, [6] reports the performance of various approaches of POS Tagging such as HMM, Maximum Entropy Model (MEM) and Conditional Random Fields (CRF) based models, as well as a memory based learning approach.

For the Tamil language, a tagger is reported in [21], which uses a suffix stripper before performing the

actual tagging, using a rule based approach to improve the accuracy.

A rule based POS tagger for Bangla is reported in [22], but only the rules for Nouns and Adjectives are showed.

Notable work on POS Tagging has been reported in [23] for Indian Bangla. Here, a HMM based approach is used for tagging.

Another paper in [24] uses a suffix based tree tagger, influenced by [18] to tag Bangla text.

A hybrid POS tagger for Bangla based on HMMs, which tags using three methods is described in [25].

We have found that most of the research on POS Tagging on the South Asian languages has been done using stochastic tagging models like HMM, MEM, CRF or some hybrid techniques as reported in [26, 27, 28, 29]. This can also be noticed in the works cited earlier.

On the other hand, a very small amount of work has been done utilizing transformation based approaches. One is described in [19]. As mentioned earlier, this is actually a hybrid tagger that primarily tags using an HMM based approach. And later the tagger retags using transformation rules. We have also tried this two phase tagging approach in the order mentioned in the paper as well as the opposite order i.e. retagging with HMMs after an initial pass using a transformation based model. But neither method improves the performance to a significant extent so as to be useful for practical purposes considering the total time and resources required for the two phases to train and tag in a sequence. And furthermore, according to the authors of that paper, the second phase sometimes reduces the accuracies for more than one tag greatly.

Other work has been found in [5], with experiments using the transformation based Brill's approach that shows very good results. This work employs Brill's approach along with some other approaches such as HMMs and n-grams on a very small corpus of around 5000 Bangla tokens, and their results indicate that transformation based approach works better for Bangla. But in this paper, the employed corpus is too small for the stochastic approaches to generate the frequency distribution tables they require to come up with appropriate tag sequences for given sentences.

4. Methodology

We used the n-grams based Unigram and Bigram, HMM based and Brill's transformation based taggers for our experiments. Details about these models can be found from various sources and search engines accessible through the internet. We used these models

provided with NLTK [30] and modified them according to our needs.

5. Corpora

We used the training, development and test data provided for the SPSAL contest [31]. We used the training data sets for each of the languages Bangla, Hindi and Telegu separately, to create our training corpora. We used the test data provided there as our testing corpora. All the data provided for the SPSAL contest uses the SSF format described in [32], which is generally used to support different kinds of linguistic analysis at different levels, such as chunking and tagging on the same data. But as we worked solely on POS Tagging for the current study, we converted all the data from the SSF format to the much simpler format used by the Brown corpus, included in NLTK [30] for our convenience.

6. Tagset

We used the 26-tags tagset provided for the SPSAL contest for our experiments with Bangla, Hindi and Telegu. The tagset is based on the Penn-Treebank tagset and consists of the following tags: Noun (NN), Proper Noun (NNP), Pronoun (PRP), Verb Finite Main (VFM), Verb Auxiliary (VAUX), Verb NonFinite Adjectival (VJJ), Verb NonFinite Adverbial (VRB), Verb NonFinite Nominal (VNN), Adjective (JJ), Adverb (RB), Noun Location (NLOC), Postposition (PREP), Particle (RP), Conjunct (CC), Question Words (QW), Quantifier (QF), Number Quantifiers (QFNUM), Intensifiers (INTF), Negative (NEG), Compound Common Nouns (NNC), Compound Proper Nouns (NNPC), Noun in kriya mula (NVB), Adj in kriya mula (JVB), Adv in kriya mula (RBVB), Interjection (UH), Special : Not classified in any other (SYM) [31].

7. Tagging examples

7.1. Untagged Bangla text

Bangla sentence:

বাড়ির বারান্দার রেলিং থেকে To Let লেখা বোর্ড ঝুলতে এখন আর কেউ দেখে কি ? তখন পাড়ায় পাড়ায় দেখা যেত ।

Bangla sentence with pronunciation:

বাড়ির/barir বারান্দার/barandār রেলিং/reliṅ থেকে/tʰeke To/tu Let/læt লেখা/lek^ha বোর্ড/bord ঝুলতে/ḷulte এখন/ek^hon

আর/ar কেউ/keu দেখে/dæk^he কি/ki ?/? তখন/tɔk^hon
পাড়ায়/paraj পাড়ায়/paraj দেখা/dæk^ha যেত/jeto.

Bangla sentence with meaning (word-to-word):

বাড়ির/(Of the) house বারান্দার/(in the) veranda
রেলিং/railing থেকে/from To/To Let/Let লেখা/written
বোর্ড/board ঝুলতে/hanging এখন/nowadays আর/and
কেউ/anyone দেখে/see কি/does ?/? তখন/At that time
পাড়ায়/(in) locality পাড়ায়/(in) locality দেখা/seeing যেত/to-
be.

Meaning of Bangla sentence in English:

Does anyone see To Let written boards hanging
from railings of verandas of houses nowadays? At that
time it was a common thing to be seen at localities.

7.2. Output of Brill's tagger on Bangla corpus using 25426 tokens.

বাড়ির/NN বারান্দার/NN রেলিং/NN থেকে/PREP To/NN
Let/NN লেখা/VFM বোর্ড/NN ঝুলতে/NN এখন/RB আর/CC
কেউ/PRP দেখে/VRB কি/QW ?/SYM তখন/RB পাড়ায়/NN
পাড়ায়/NN দেখা/VFM যেত/VAUX । /SYM

7.3. Output of Unigram tagger on Bangla corpus using 25426 tokens.

বাড়ির/NNP বারান্দার/NNP রেলিং/NNP থেকে/PREP To/NNP
Let/NNP লেখা/VFM বোর্ড/NNP ঝুলতে/NNP এখন/RB
আর/CC কেউ/PRP দেখে/VRB কি/QW ?/SYM তখন/RB
পাড়ায়/NNP পাড়ায়/NNP দেখা/VFM যেত/VAUX । /SYM

7.4. Output of Bigram tagger on Bangla corpus using 25426 tokens.

বাড়ির/NNP বারান্দার/NNP রেলিং/NNP থেকে/PREP To/NNP
Let/NNP লেখা/VFM বোর্ড/NNP ঝুলতে/NNP এখন/RB
আর/CC কেউ/PRP দেখে/VRB কি/QW ?/SYM তখন/CC
পাড়ায়/NNP পাড়ায়/NNP দেখা/VFM যেত/VAUX । /SYM

7.5. Output of HMM tagger on Bangla corpus using 25426 tokens.

বাড়ির/NNPC বারান্দার/NNPC রেলিং/NNP থেকে/PREP
To/QFNUM Let/NN লেখা/VFM বোর্ড/VAUX ঝুলতে/SYM
এখন/PRP আর/CC কেউ/PRP দেখে/VRB কি/QW ?/SYM
তখন/PRP পাড়ায়/JJ পাড়ায়/NN দেখা/VFM যেত/VAUX
। /SYM

7.6. Untagged Telegu text

nenemayyAnA , yeVMxuku rAleraxA anu_kuMtU
amma , nAnnA nA kosaM kanipeVttukunnAru .

7.7. Output of Brill's tagger on Telegu corpus using 27511 tokens.

nenemayyAnA/NN ,/SYM yeVMxuku/NN
rAleraxA/NN anu_kuMtU/NN amma/NN ,/SYM
nAnnA/NN nA/PRP kosaM/PREP
kanipeVttukunnAru/VFM ./SYM

7.8. Output of Unigram tagger on Telegu corpus using 27511 tokens.

nenemayyAnA/NNP ,/SYM yeVMxuku/NNP
rAleraxA/NNP anu_kuMtU/NNP amma/NNP ,/SYM
nAnnA/NNP nA/PRP kosaM/PREP
kanipeVttukunnAru/NNP ./SYM

7.9. Output of Bigram tagger on Telegu corpus using 27511 tokens.

nenemayyAnA/NNP ,/SYM yeVMxuku/NNP
rAleraxA/NNP anu_kuMtU/NNP amma/NNP ,/SYM
nAnnA/NNP nA/PRP kosaM/PREP
kanipeVttukunnAru/NNP ./SYM

7.10. Output of HMM tagger on Telegu corpus using 27511 tokens.

nenemayyAnA/QFNUM ,/SYM yeVMxuku/QW
rAleraxA/VFM anu_kuMtU/SYM amma/NN ,/SYM
nAnnA/CC nA/PRP kosaM/PREP
kanipeVttukunnAru/VFM ./SYM

8. Results

We experimented with Unigram, Bigram, HMM
and Brill's tagger on Bangla, Hindi and Telegu. For
Bangla we used a training corpus with a maximum of
1786 sentences consisting of 25426 tokens. Our test
corpus consisted of 400 sentences and 5225 tokens.
The performances of the taggers are shown below:

**Table 1. Performance of taggers on the Bangla
corpus.**

HMM	Unigram	Bigram	Brill
63.6	56.9	55.5	69.6

For Hindi we used a training corpus with a
maximum of 1135 sentences consisting of 26148

tokens. Our test corpus consisted of 209 sentences and 4924 tokens. The performance of the taggers are shown below:

Table 2. Performance of taggers on the Hindi corpus.

HMM	Unigram	Bigram	Brill
68.5	58.5	57.5	71.5

For Telegu we used a training corpus with a maximum of 2655 sentences consisting of 27511 tokens. Our test corpus consisted of 415 sentences and 5193 tokens. The performance of the taggers are shown below:

Table 3. Performance of taggers on the Telegu corpus.

HMM	Unigram	Bigram	Brill
56.6	42.8	42.2	66.9

The graphs denoting the changes in performance of the taggers with corpora size is included below.

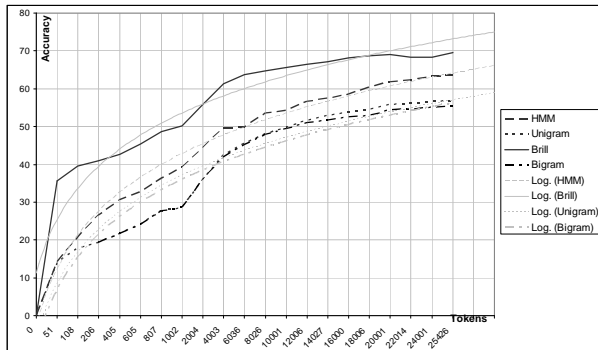


Figure 2. Corpora size vs. performance of taggers on Bangla corpus.

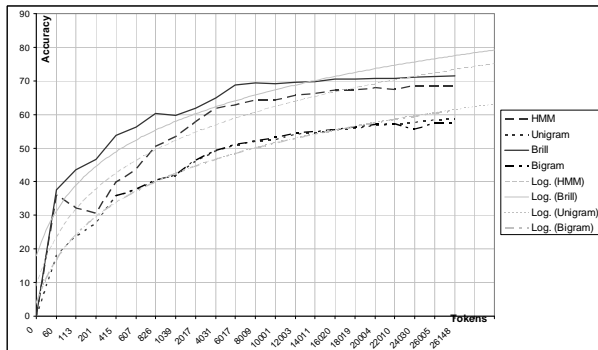


Figure 3. Corpora size vs. performance of taggers on Hindi corpus.

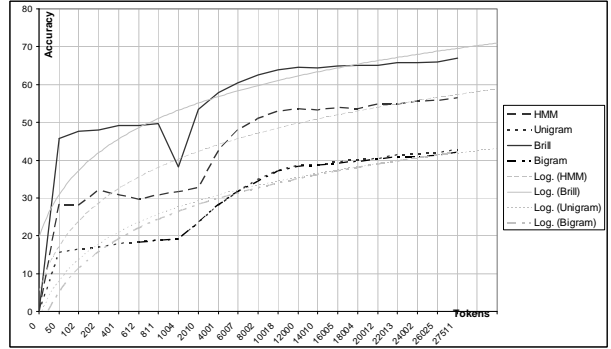


Figure 4. Corpora size vs. performance of taggers on Telegu corpus.

In all the above cases the line at the top denoting better performance is the Brill tagger. The line in the middle is that of the HMM tagger. The Unigram and Bigram taggers are seen at the bottom and are very close to each other.

9. Analysis of results

We have observed from the results of a previous study described in [5], that the HMM based tagger performs better than transformation based or n-grams based taggers starting from a very small corpus for English using the Brown corpus provided in NLTK [30]. The difference in performance also continues to grow as the corpus size increases. We have also found that according to this study, the opposite happens for Bangla using a very small corpus of around 4000 tokens. That is, Brill's tagger performs better than the HMM based model for Bangla.

In our present work, we used corpora with over 25000 annotated tokens for each of the languages Bangla, Hindi and Telegu. We used a different tagset for Bangla than that mentioned in [5].

Under these conditions, we observed that Brill's tagger achieves accuracies of 69.6% using 25226 tokens for Bangla, 71.5% using 26148 tokens on Hindi and 66.9% using 27511 tokens on Telegu, whereas the HMM tagger manages to obtain 63.6%, 68.5% and 56.6%. The Unigram and Bigram taggers manage 56.9%, 58.5% and 42.8%; and 55.5%, 57.5% and 42.2% respectively, using the same number of tokens as Brill's tagger. These results are also comparable and fall in the same range as those of the SPSAL contest [31].

It can be noticed from the results that Brill's tagger not only performs better than other taggers for Bangla, but it also outperforms other taggers significantly for Hindi and Telegu as well. So the experiments confirm

that Brill's tagger is a better choice for tagging South Asian Language using small to medium sized corpora.

The reason behind the superior performance of the Brill tagger could be the structure of the three languages we experimented with.

South Asian languages are generally rule based and can be described in rules as they descend from the same root language, Sanskrit, which itself strictly maintains grammatical rules and can almost entirely be described with rules.

And as we know from [11, 12], Brill's transformation based tagger is actually an enhanced type of rule based tagger that uses pre-written rules and induces new rules while training, so it would be quite natural for it to perform better in tagging rule based South Asian languages than taggers based on different models.

10. Future work

Several modifications to the baseline POS taggers are described in [15, 33, 34] that suggest the use of techniques like pre-tagging problematic idioms, using Finite State Transducers (FST) to speed up the operation of the tagger. We would like to incorporate these in our tagging models.

The transformation based Brill's approach can be used unsupervised as described in [12]. We have left the unsupervised approaches out of the scope of the present study, because of their high requirements of computational power and slow speed to train. But for South Asian languages, in which Brill's tagger has good prospects, and large training corpora may not be readily available, the unsupervised or semi-supervised transformation based approach could prove very useful. We would like to experiment with that model in our next studies.

From [35] we have come to know of three patterns of behavior in Baum Welch re-estimation. The next step could be to find out whether these patterns are present in South Asian Languages, and also, whether the guidelines mentioned in the paper are applicable for these languages as well.

In the previous section, we developed a proposition about the superior performance of Brill's tagger for South Asian languages. But more experiments need to be done on this before reaching a conclusion. We would continue to work on the subject to check whether the proposition holds true under different conditions.

11. Conclusion

We compared the performance of n-grams, HMM or transformation based POS Taggers on three South Asian Languages, Bangla, Hindi and Telegu. And we found that the HMM based tagger might perform better for English, but for South Asian languages, using corpora of different sizes, the transformation based Brill's approach performs significantly better than any other approach when using a 26-tags tagset and pre-annotated training corpora consisting of a maximum of 25426, 26148 and 27511 tokens for Bangla, Hindi and Telegu respectively. We also proposed a reason behind the better performance of the transformation based approach. So researchers working on these languages might try out the transformation based approach alongside the other widely used approaches.

12. Acknowledgement

This work has been supported in part by the PAN Localization Project (www.panl10n.net) grant from the International Development Research Center, Ottawa, Canada, administrated through the Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan.

13. References

- [1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, Prentice Hall, 2000.
- [2] Andrew MacKinlay, "The Effects of Part-of-Speech Tagsets on Tagger Performance", Undergraduate Thesis, University of Melbourne, 2005.
- [3] Yair Halevi, "Part of Speech Tagging", *Seminar in Natural Language Processing and Computational Linguistics (Prof. Nachum Dershowitz)*, School of Computer Science, Tel Aviv University, Israel, April 2006.
- [4] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.
- [5] Fahim Muhammad Hasan, Naushad UzZaman, Mumit Khan, "Comparison of Different POS Tagging Techniques (n-grams, HMM and Brill's Tagger) for Bangla", *International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06)*, December 4-14, 2006.

- [6] Karthik Kumar G, Sudheer K, Avinesh Pvs, "Comparative Study of Various Machine Learning Methods for Telugu Part of Speech Tagging", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [7] Linda Van Guilder, "Automated Part of Speech Tagging: A Brief Overview", Handout for LING361, Georgetown University, Fall 1995.
- [8] Manoj Kumar C, "Stochastic Models for POS Tagging", IIT Bombay, 2005.
- [9] L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation on Probabilistic Functions of a Markov Process", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 22, Issue: 4, April 2000, pp. 371-377.
- [10] Juan Antonio P'erez-Ortiz and Mikel L. Forcada, "Part-of-Speech Tagging with Recurrent Neural Networks", Universitat d'Alacant, Spain, 2002.
- [11] Eric Brill, "A Simple Rule-Based Part-of-Speech Tagger", *In Proceeding Of The Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992, pp. 152-155.
- [12] Eric Brill, "Some Advances in Transformation Based Part of Speech Tagging", *In Proceeding of The Twelfth National Conference on Artificial Intelligence (vol. 1)*, Seattle, Washington, United States, 1994, pp. 722-727.
- [13] L. Bahl and R. L. Mercer, "Part-Of-Speech Assignment by a Statistical Decision Algorithm", *IEEE International Symposium on Information Theory*, 1976, pp. 88-89.
- [14] K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser For Unrestricted Test", *Proceeding of the Second Conference on Applied Natural Language Processing*, 1988, pp. 136-143.
- [15] D. Cutting, J. Kupiec, J. Pederson and P. Sibun, "A Practical Part-Of-Speech Tagger", *Proceeding of the Third Conference on Applied Natural Language Processing, ACL*, Trento, Italy, 1992, pp. 133-140.
- [16] S. J. DeRose, "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics*, 14 (1), 1988.
- [17] A. M. Deroualt and B. Merialdo, "Natural Language Modeling For Phoneme-To-Text Transposition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.
- [18] Helmut Schmidt, "Improvements in POS Tagging with an Application to German", *Technical Report*, Universitat Stuttgart, Germany, 2005.
- [19] Pranjal Awasthi, Delip Rao and Balaraman Ravindran, "Part Of Speech Tagging and Chunking with HMM and CRF", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [20] Smriti Singh, Kuhoo Gupta, Manish Shrivastava and Pushpak Bhattacharyya, "Morphological Richness Offsets Resource Demand- Experiences in Constructing a POS Tagger for Hindi", *In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, July 2006, pp. 779-786.
- [21] Arulmozhi. P, Sobha. L, Kumara Shanmugam. B., "Parts of Speech Tagger for Tamil", *Symposium on Indian Morphology, Phonology and Language Engineering*, IIT Khadagpur, India, March 19-21, 2004, pp. 55-57.
- [22] Md. Shahnur Azad Chowdhury, Nahid Mohammad Minhaz Uddin, Mohammad Imran, Mohammad Mahadi Hassan, and Md. Emdadul Haque, "Parts of Speech Tagging of Bangla Sentence", *In Proceeding of the 7th International Conference on Computer and Information Technology (ICCIIT)*, Bangladesh, 2004.
- [23] Sandipan Dandapat, Sudeshna Sarkar, "Part of Speech Tagging for Bengali with Hidden Markov Model", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [24] Md. Hanif Seddiqui, A. K. Muhammad Shohel Rana, Abdullah Al Mahmud and Taufique Sayeed, "Parts of Speech Tagging Using Morphological Analysis in Bangla", *In Proceeding of the 6th International Conference on Computer and Information Technology (ICCIT)*, Bangladesh, 2003.
- [25] Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu, "A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali", *International Journal of Information Technology Volume 1 Number 4*, 2004.
- [26] Sankaran Baskaran, "Hindi POS Tagging and Chunking", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [27] Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, "Hindi Part-of-Speech Tagging and Chunking : A Maximum Entropy Approach", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [28] Sivaji Bandyopadhyay, Asif Ekbal and Debasish Halder, "HMM Based POS Tagger and Rule-Based Chunker for Bengali", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.
- [29] Himanshu Agrawal and Anirudh Mani, "Part of Speech Tagging and Chunking with Conditional Random Fields", *In Proceeding of the NLP AI Machine Learning Competition*, 2006.

[30] Steven Bird and Edward Loper, "Natural Language Toolkit", <http://nltk.sourceforge.net/>, 2006.

[31] Workshop on Shallow Parsing in South Asian Languages (SPSAL) 2007, *Twentieth International Joint Conferences on Artificial Intelligence*, Hyderabad, India.

[32] Akshar Bharati, Rejeev Sangal and Dipti M Sharma, "Shakti Analyser: SSF Representation", IIT Hyderabad, 2006.

[33] Emmanuel Roche and Yves Schabes, "Deterministic POS Tagging with Finite State Transducers (FST)", *Computational Linguistics*, Volume 21, Issue 2, June 1995, pp. 227-253.

[34] Virginia Savova and Leonid Peshkin, "Part-of-Speech Tagging with Minimal Lexicalization", Johns Hopkins University, Massachusetts Institute of Technology.

[35] David Elworthy, "Does Baum-Welch Re-estimation Help Taggers?", *In Proceeding of The Fourth Conference on Applied Natural Language Processing*, 1994, pp. 53-58.