Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla

Fahim Muhammad Hasan, Naushad UzZaman and Mumit Khan
Center for Research on Bangla Language Processing, BRAC University, Bangladesh
stealth_310@yahoo.com, naushad@bracuniversity.net, mumit@bracuniversity.net

Abstract

There are different approaches to the problem of assigning each word of a text with a parts-of-speech tag, which is known as Part-Of-Speech (POS) tagging. In this paper we compare the performance of a few POS tagging techniques for Bangla language, e.g. statistical approach (n-gram, HMM) and transformation based approach (Brill's tagger). A supervised POS tagging approach requires a large amount of annotated training corpus to tag properly. At this initial stage of POS-tagging for Bangla, we have very limited resource of annotated corpus. We tried to see which technique maximizes the performance with this limited resource. We also checked the performance for English and tried to conclude how these techniques might perform if we can manage a substantial amount of annotated corpus.

1. Introduction

Bangla is among the top ten most widely spoken languages [1] with more than 200 million native speakers, but it still lacks significant research efforts in the area of natural language processing.

Part-of-Speech (POS) tagging is a technique for assigning each word of a text with an appropriate parts of speech tag. The significance of part-of-speech (also known as POS, word classes, morphological classes, or lexical tags) for language processing is the large amount of information they give about a word and its neighbor. POS tagging can be used in TTS (Text to Speech), information retrieval, shallow parsing, information extraction, linguistic research for corpora [2] and also as an intermediate step for higher level NLP tasks such as parsing, semantics, translation, and many more [3]. POS tagging, thus, is a necessary application for advanced NLP applications in Bangla or any other languages.

We start this paper by giving an overview of a few POS tagging models; we then discuss what have been done for Bangla. Then we show the methodologies we used for POS tagging; then we describe our POS tagset, training and test corpus; next we show how these methodologies perform for both English and Bangla; finally we conclude how Bangla (language with limited language resources, tagged corpus) might perform in comparison to English (language with available tagged corpus).

2. Literature review

Different approaches have been used for Part-of-Speech (POS) tagging, where the notable ones are rule-based, stochastic, or transformation-based learning approaches. Rule-based taggers [4, 5, 6] try to assign a tag to each word using a set of hand-written rules. These rules could specify, for instance, that a word following a determiner and an adjective must be a noun. Of course, this means that the set of rules must be properly written and checked by human experts. The stochastic (probabilistic) approach [7, 8, 9, 10] uses a training corpus to pick the most probable tag for a word. All probabilistic methods cited above are based on first order or second order Markov models. There are a few other techniques which use probabilistic approach for POS Tagging, such as the Tree Tagger [11]. Finally, the transformation-based approach combines the rule-based approach and statistical approach. It picks the most likely tag based on a training corpus and then applies a certain set of rules to see whether the tag should be changed to anything else. It saves any new rules that it has learnt in the process, for future use. One example of an effective tagger in this category is the Brill tagger [12, 13, 14, 15].

All of the approaches discussed above fall under the rubric of supervised POS Tagging, where a pre-tagged corpus is a prerequisite. On the other hand, there is the unsupervised POS tagging [16, 17, 18] technique, and it does not require any pre-tagged corpora.

Figure 1 demonstrates the classification of different POS tagging schemes.

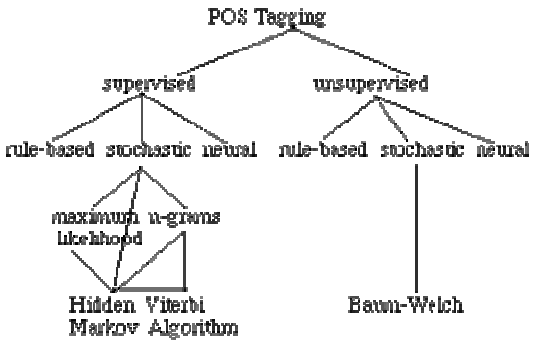

Figure 1: Classification of POS tagging models [19]

For English and many other western languages many such POS tagging techniques have been implemented and in almost all the cases, they show a satisfying performance of 96+%. For Bangla work on POS tagging has been reported by [20, Chowdhury et al. (2004) and Seddiqui et al. (2003).

Chowdhury et al. (2004) implemented a rule based POS tagger, which requires writing laboriously handcrafted rules by human experts and many years of continuous efforts from many linguists. Since they report no performance analysis of their work, the feasibility of their proposed rule based method for Bangla is suspect. No review or comparison of established work on Bangla POS tagging was available in that paper; they only proposed a rule-based technique. Their work can be described as more of a morphological analyzer than a POS tagger. A morphological analyzer indeed provides some POS tag information, but a POS-tagger needs to operate on a large set of fine-grained tags. For example, the [23] for English consists of 87 distinct tags, and Penn Treebank's [24] tagset consists of 48 tags. Chowdhury et al.'s tagset, by contrast, consists of only 9 tags and they showed only rules for nouns and adjectives for their POS Tagger. Such a POS-tagger's output will have very limited applicability in many advanced NLP applications.

For English, researchers had tried this rule-based technique in the 60s and 70s [4, 5, 6]. Taking into consideration of the problem of this method, researchers have switched to statistical or machine learning methods, or more recently, to the unsupervised methods for POS tagging.

In this paper we compare the performance of different tagging techniques such as Brill's tagger, n-gram tagger and HMM tagger for Bangla; such comparison was not attempted in [20, 21, 22].

3. Methodology

NLTK [25], the Natural Language Toolkit, is a suite of program modules, data sets and tutorials supporting research and teaching in computational linguistics and natural language processing. NLTK has many modules implemented for different NLP applications. We have experimented unigram, bigram, HMM and Brill tagging modules from NLTK [25] for our purpose.

3.1. Unigram tagger

The unigram (n-gram, n = 1) tagger is a simple statistical tagging algorithm. For each token, it assigns the tag that is most likely for that token's text. For example, it will assign the tag `jj` to any occurrence of the word `frequent`, since `frequent` is used as an adjective (e.g. a `frequent` word) more often than it is used as a verb (e.g. I `frequent` this cafe).

Before a unigram tagger can be used to tag data, it must be trained on a training corpus. It uses the corpus to determine which tags are most common for each word.

The unigram tagger will assign the default tag `None` to any token that was not encountered in the training data.

3.2. HMM

The intuition behind HMM (Hidden Markov Model) and all stochastic taggers is a simple generalization of the "pick the most likely tag for this word" approach. The unigram tagger only considers the probability of a word for a given tag *t*; the surrounding context of that word is not considered.

On the other hand, for a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the following formula:

`P (word | tag) * P (tag | previous` *n* `tags)`

3.3. Brill's transformation based tagger

A potential issue with nth-order tagger is their size. If tagging is to be employed in a variety of language technologies deployed on mobile computing devices, it is important to find ways to reduce the size of models without overly compromising performance. An nth-order tagger with backoff may store trigram and bigram tables, large sparse arrays, which may have hundreds of millions of entries. A consequence of the size of the models is that it is simply impractical for

nth-order models to be conditioned on the identities of words in the context. In this section we will examine Brill tagging, a statistical tagging method which performs very well, using models that are only a tiny fraction of the size of nth-order taggers.

Brill tagging is a kind of transformation-based learning. The general idea is very simple: guess the tag of each word, then go back and fix the mistakes. In this way, a Brill tagger successively transforms a bad tagging of a text into a good one. As with nth-order tagging this is a supervised learning method, since we need annotated training data. However, unlike nth-order tagging, it does not count observations but compiles a list of transformational correction rules.

The process of Brill tagging is usually explained by analogy with painting. Suppose we were painting a tree, with all its details of boughs, branches, twigs and leaves, against a uniform sky-blue background. Instead of painting the tree first then trying to paint blue in the gaps, it is simpler to paint the whole canvas blue, then "correct" the tree section by overpainting the blue background.

In the same fashion we might paint the trunk a uniform brown before going back to overpaint further details with a fine brush. Brill tagging uses the same idea: get the bulk of the painting right with broad brush strokes, then fix up the details. As time goes on, successively finer brushes are used, and the scale of the changes becomes arbitrarily small. The decision of when to stop is somewhat arbitrary.

In our experiment we have used the taggers (Unigram, HMM, Brill's transformation based tagger) described above. Detailed descriptions of these taggers are available at [2, 26].

4. POS tagset

For English we have used the Brown Tagset [23]. And for Bangla we have used a 41 tag-sized tagset [28]. Our tagset has two levels of tags. First level is the high-level tag for Bangla, which consists of only 12 tags (Noun, Adjective, Cardinal, Ordinal, Fractional, Pronoun, Indeclinable, Verb, Post Positions, Quantifiers, Adverb, Punctuation). And the second level is more fine-grained with 41 tags. Most of our experiments are based on the level 2 tagset (41 tags). However, we experimented few cases with level 1 tagset (12 tags).

5. Training corpus and test set

For our experiment for English, we have used tagged Brown corpus from NLTK [25]. For Bangla, we have a very small corpus of around 5000 words from a Bangladeshi daily newspaper Prothom-alo [27]. In both cases, our test set is disjoint from the training corpus.

6. Tagging example

Bangla (Training corpus size: 4484 tokens)

Untagged Text:

1. দ্বিতীয় বিশ্বযুদ্ধে মিত্র বাহিনীর নেতা ব্রিটিশ প্রধানমন্ত্রী উইন্সটন চার্চিলকে গত সপ্তাহের শুরুতে টপকে বেয়ার এ তালিকায় স্থান লাভ করেন ।
2. তবে তিনি যদি আবার নির্বাচন করেন এবং জয়ী হন তাহলে হয়তো এ রেকর্ডও ভাঙতে পারবেন ।

Tagged output:

Level 2 Tagset (41 Tags)

Brill:

1. দ্বিতীয়/NC বিশ্বযুদ্ধে/NC মিত্র/NC বাহিনীর/NC নেতা/NC ব্রিটিশ/ADJ প্রধানমন্ত্রী/NC উইন্সটন/NP চার্চিলকে/NP গত/ADJ সপ্তাহের/NC শুরুতে/ADVT টপকে/NP বেয়ার/NP এ/DP তালিকায়/NC স্থান/NC লাভ/NC করেন/VF ।/PUNSF
2. তবে/INDO তিনি/PP যদি/INDO আবার/ADVM নির্বাচন/NC করেন/VF এবং/CONJC জয়ী/NC হন/VE তাহলে/INDO হয়তো/OTHER এ/DP রেকর্ডও/NC ভাঙতে/NC পারবেন/VF ।/PUNSF

Unigram:

1. দ্বিতীয়/NP বিশ্বযুদ্ধে/NP মিত্র/NP বাহিনীর/NC নেতা/NC ব্রিটিশ/ADJ প্রধানমন্ত্রী/NC উইন্সটন/NP চার্চিলকে/NP গত/ADJ সপ্তাহের/NC শুরুতে/ADVT টপকে/NP বেয়ার/NP এ/DP তালিকায়/NC স্থান/NC লাভ/NP করেন/VF ।/PUNSF
2. তবে/INDO তিনি/PP যদি/INDO আবার/ADVM নির্বাচন/NC করেন/VF এবং/CONJC জয়ী/NP হন/VE তাহলে/INDO হয়তো/OTHER এ/DP রেকর্ডও/NP ভাঙতে/NP পারবেন/NP ।/PUNSF

HMM:

1. দ্বিতীয়/DP বিশ্বযুদ্ধে/NC মিত্র/NC বাহিনীর/NC নেতা/NC ব্রিটিশ/ADJ প্রধানমন্ত্রী/NC উইন্সটন/NP চার্চিলকে/NP গত/ADJ সপ্তাহের/NC শুরুতে/ADVT টপকে/ADVT বেয়ার/NP এ/NP তালিকায়/NC স্থান/NC লাভ/NC করেন/VF ।/PUNSF
2. তবে/INDO তিনি/PP যদি/INDO আবার/ADVM নির্বাচন/NC করেন/VF এবং/CONJC জয়ী/NC হন/VF তাহলে/PUNSF হয়তো/OTHER এ/DP রেকর্ডও/NC ভাঙতে/VNF পারবেন/VF ।/PUNSF

Level 1 Tagset (Reduced Tagset: 12 Tags)

Brill:

1. দ্বিতীয়/NN বিশ্বযুদ্ধে/NN মিত্র/NN বাহিনীর/NN নেতা/NN ব্রিটিশ/ADJ প্রধানমন্ত্রী/NN উইন্সটন/NN চার্চিলকে/NN গত/ADJ সপ্তাহের/NN শুরুতে/ADV টপকে/NN বেয়ার/NN এ/PN তালিকায়/NN স্থান/NN লাভ/NN করেন/VB ।/PUNC
2. তবে/IND তিনি/PN যদি/IND আবার/ADV নির্বাচন/NN করেন/VB এবং/IND জয়ী/NN হন/VB তাহলে/IND হয়তো/OTHER এ/PN রেকর্ডও/NN ভাঙতে/VB পারবেন/VB ।/PUNC

Unigram:

1. দ্বিতীয়/NN বিশ্বযুদ্ধে/NN মিত্র/NN বাহিনীর/NN নেতা/NN ব্রিটিশ/ADJ প্রধানমন্ত্রী/NN উইন্সটন/NN চার্চিলকে/NN গত/ADJ সপ্তাহের/NN শুরুতে/ADV টপকে/NN বেয়ার/NN এ/PN তালিকায়/NN স্থান/NN লাভ/NN করেন/VB ।/PUNC
2. তবে/IND তিনি/PN যদি/IND আবার/ADV নির্বাচন/NN করেন/VB এবং/IND জয়ী/NN হন/VB তাহলে/IND হয়তো/OTHER এ/PN রেকর্ডও/NN ভাঙতে/VB পারবেন/NN ।/PUNC

HMM:

1. দ্বিতীয়/PN বিশ্বযুদ্ধে/NN মিত্র/NN বাহিনীর/NN নেতা/NN ব্রিটিশ/ADJ প্রধানমন্ত্রী/NN উইন্সটন/NN চার্চিলকে/NN গত/ADJ সপ্তাহের/NN শুরুতে/ADV টপকে/ADV বেয়ার/NN এ/NN তালিকায়/NN স্থান/NN লাভ/NN করেন/VB ।/PUNC
2. তবে/IND তিনি/PN যদি/IND আবার/ADV নির্বাচন/NN করেন/VB এবং/IND জয়ী/NN হন/VB তাহলে/IND হয়তো/OTHER এ/PN রেকর্ডও/NN ভাঙতে/VB পারবেন/VB ।/PUNC

7. Performance

We have experimented POS taggers (Unigram, HMM, Brill) for both Bangla and English. For Bangla we experimented in both tag levels (level 1 – 12 tags, level 2 – 41 tags). Experiment results are given below in form of table and graph.

Table 1: Performance of POS Taggers for Bangla [Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus; Tagset: Level 1 Tagset (12 Tags)]

Tokens	HMM Accuracy	Unigram Accuracy	Brill Accuracy
0	0	0	0
60	15.4	51.2	50.4
104	18	51.1	44.6
503	34.2	60.7	56.3
1011	42.3	64.2	62.6
2023	45.8	69.1	67.8
3016	49.4	70.1	70.9
4484	45.6	71.2	71.3

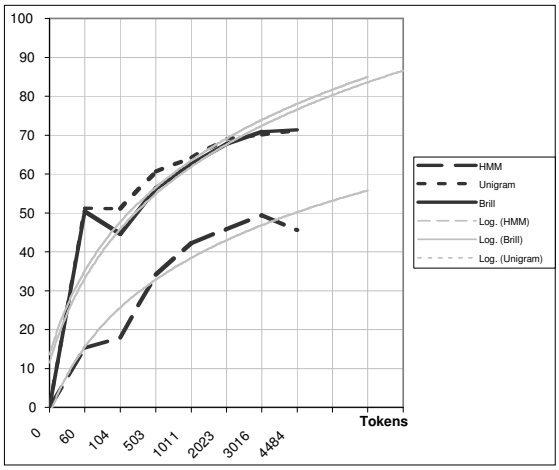

Figure 1: Performance of POS Taggers for Bangla [Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus; Tagset: Level 1 Tagset (12 Tags)]

Table 2: Performance of POS Taggers for Bangla [Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus; Tagset: Level 2 Tagset (41 Tags)]

Tokens	HMM Accuracy	Unigram Accuracy	Brill Accuracy
0	0	0	0
60	19.7	17.2	38.7
104	18.1	17.4	26.2
503	28.8	26.1	46.1
1011	32.8	30	51.1
2023	40.1	36.7	49.4
3016	44.5	39.1	51.9
4484	46.9	42.2	54.9

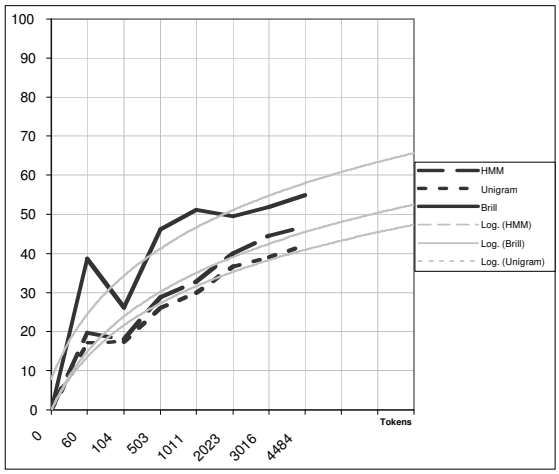

Figure 2: Performance of POS Taggers for Bangla [Test data: 85 sentences, 1000 tokens from the (Prothom-Alo) corpus; Tagset: Level 2 Tagset (41 Tags)]

Table 3: Performance of POS Taggers for English [Test data: 22 sentences, 1008 tokens from the Brown corpus; Tagset: Brown Tagset]

Tokens	HMM Accuracy	Unigram Accuracy	Brill Accuracy
0	0	0	0
65	36.9	28.7	33.6
134	44.2	34	42.9
523	53.4	41.6	53.7
1006	62	47.7	58.3
2007	66.8	52.4	62.9
3003	68.2	55.1	66.1
4042	70	57.2	67.5
5032	71.5	59.2	70.2
6008	71.9	60.8	71.4
7032	74.5	61.5	71.8
8010	74.8	62.1	72.4
9029	76.8	63.5	74.5
10006	77.5	65.2	75.2
20011	80.9	69.5	79.8
30017	83.1	71.7	78.8
40044	84.7	73.3	79.8
50001	84.6	74.4	80.4
60022	85.3	75.2	80.8
70026	86.3	75.8	81
80036	87.1	77.1	81.6
90000	87.8	78.1	82.4
100057	87.5	78.9	83.4
200043	91.7	83	86.8
300359	89.5	84.2	87.3
400017	89.7	84.8	88.5
500049	90.3	85.6	
600070	90	85.9	
700119	90.3	86.1	
800031	90.2	86.2	
900073	90.3	86.6	
1000107	90.3	86.5	

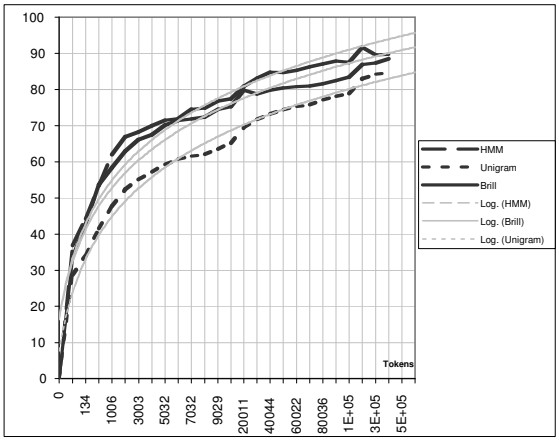

Figure 3: Performance of POS Taggers for English [Test data: 22 sentences, 1008 tokens from the Brown corpus; Tagset: Brown Tagset]

8. Analysis of test result

English POS taggers report high accuracy of 96+%, where the same taggers did not perform the same (only 90%) in our case. This is because others tested on a large training set for their taggers, whereas we tested our English taggers on a maximum of 1 million sized corpus (for HMM and unigram) and for Brill, we tested under training of 400 thousand tokens.

Since our Bangla taggers were being tested on a very small-sized corpus (with a maximum of 4048 tokens), the resulting performance by them was not satisfactory. This was expected, however, as the same taggers performed similarly for a similar-sized English corpus (see Table 3). For English we have seen that performance increases with the increase of corpus size. For Bangla we have seen it follows the same trend as English. So, it can be safely hypothesized that if we can extend the corpus size of Bangla then we will be able to get the similar performance for Bangla as English.

Within this limited corpus (4048 tokens), our experiment suggests that for Bangla (both with 12-tag tagset and 41-tag tagset), Brill's tagger performed better than HMM-based tagger and Unigram tagger (see Tables 1, 2). Researchers who are studying a sister language of Bangla and want to implement a POS tagger can try Brill's tagger, at least for a small-sized corpus.

9. Future work

Unsupervised POS tagging is a very good choice for languages with limited POS tagged corpora. We want to check how Bangla performs using unsupervised POS tagging techniques.

In parallel to the study of unsupervised techniques, we want to try a few other state of the art POS tagging techniques for Bangla. In another study we have seen that in case of n-gram based POS tagging, backward n-gram (considers next words) performs better than usual forward n-gram (considers previous words).

Our final target is to propose a hybrid solution for POS tagging in Bangla that performs with 95%+ as in English or other western languages and use this POS tagger in other advanced NLP applications.

10. Conclusion

We showed that using n-gram (unigram), HMM and Brill's transformation based techniques, the POS tagging performance for Bangla is approaching that of English. With the training set of around 5000 words and a 41-tag tagset, we get a performance of 55%. With a much larger training set, it should be possible to increase the level of accuracy of Bangla POS taggers comparable to the one achieved by English POS taggers.

11. 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 Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan.

12. References

[1] The Summer Institute for Linguistics (SIL) Ethnologue Survey, 1999.

[2] D. Jurafsky and J.H. Martin, "Chapter 8: Word classes and Part-Of-Speech Tagging", *Speech and Language Processing*, Prentice Hall, 2000.

[3] Y. 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] B. Greene and G. Rubin, "Automatic Grammatical Tagging of English", *Technical Report, Department of Linguistics, Brown University*, Providence, Rhode Island, 1971.

[5] S. Klein and R. Simmons, "A computational approach to grammatical coding of English words", *JACM 10*, 1963.

[6] Z. Harris, *String Analysis of Language Structure*, Mouton and Co., The Hague, 1962.

[7] 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.

[8] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted test", In proceeding of the *Second Conference on Applied Natural Language Processing*, 1988, pp. 136 - 143.

[9] D. Cutting, J. Kupiec, J. Pederson and P. Sibun, "A practical Part-Of-Speech Tagger", In *proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, Italy, 1992, pp. 133 - 140.

[10] S. J. DeRose, "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics, 14 (1)*, 1988.

[11] H. Schmid, "Probabilistic Part-Of-Speech Tagging using Decision Trees", In *Proceedings of the International Conference on new methods in language processing*, Manchester, UK, 1994, pp. 44-49.

[12] E. Brill, "A simple rule based part of speech tagger", In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, Trento, Italy, 1992.

[13] E. Brill, "Automatic grammar induction and parsing free text: A transformation based approach",

In *proceedings of 31st Meeting of the Association of Computational Linguistics*, Columbus, Oh, 1993.

[14] E. Brill, "Transformation based error driven parsing", In *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands, 1993.

[15] E. Brill, "Some advances in rule based part of speech tagging", In *Proceedings of The Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Washington, 1994.

[16] R. Prins and G. van Noord, "Unsupervised Pos-Tagging Improves Parsing Accuracy And Parsing Efficiency", In *Proceedings of the International Workshop on Parsing Technologies*, 2001.

[17] M. Pop, "Unsupervised Part-of-speech Tagging", Department of Computer Science, Johns Hopkins University, 1996.

[18] E. Brill, "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging", In *Proceeding of The Natural Language Processing Using Very Large Corpora*, Boston, MA, 1997.

[19] L. van Guilder, "Automated Part of Speech Tagging: A Brief Overview", *Handout for LING361*, Fall 1995, Georgetown University.

[20] S. Dandapat, S. Sarkar and A. Basu, "A Hybrid Model for Part-Of-Speech Tagging and its Application to Bengali", In *Proceedings of the International Journal of Information Technology, Volume 1, Number 4*.

[21] M.S.A. Chowdhury, N.M. Minhaz Uddin, M. Imran, M.M. Hassan, and M.E. Haque, "Parts of Speech Tagging of Bangla Sentence", In *Proceeding of the 7th International Conference on Computer and Information Technology (ICCIT)*, Bangladesh, 2004.

[22] M.H. Seddiqui, A.K.M.S. Rana, A. Al Mahmud and T. 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.

[23] Brown Tagset, available online at: http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html

[24] M.P. Marcus, B. Santorini and M.A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics Journal, Volume 19, Number 2, 1994, pp. 313 - 330. Available online at: http://www.ldc.upenn.edu/Catalog/docs/treebank2/cl93.html

[25] NLTK, The Natural Language Toolkit, available online at: http://nltk.sourceforge.net/index.html

[26] NLTK's tagger documentation, available online at: http://nltk.sourceforge.net/tutorial/tagging.pdf

[27] Bangla Newspaper, Prothom-Alo. Online version available online at: http://www.prothom-alo.net

[28] Bangla POS Tagset used in our Bangla POS tagger, available online at http://www.naushadzaman.com/bangla_tagset.pdf