**Abstract**

Parts of speech tagging is widely used for natural language processing.In this paper we describe a simple corpus-based and rule-based approach for bangla parts of speech tagger.In our parts of speech(POS) tagger we used bangla word stemming.We defined some rules for stemming bangla words.We have also proposed chain rule for tagging unknown words.

**Introduction**

Parts of speech tagging[1] is a process where each word in a sentence is tagged with appropriate parts of speech for that word. This is also called as grammatical tag. Based on both its definition, as well as its context, the process of marking up the words in a text (corpus) as corresponding to a particular part of speech is named as parts of speech tagging or POS tagging. It may be called as the identification of a word in a sentence. Parts of speech tagging is widely used for linguistic text analysis. Part-of-speech tagging is an essential task for all the natural language processing activities. Parts of speech tagging is harder because of different parts of speech for a single word when we try to do it automatically, it is much harder when we try to do for a language like bangla, where the bangla language has less contextual and syntactical clue and need much morphological analysis.

**Common approaches of POS tagging**

A lot of work has been done for parts of speech tagging. There are several

procedures for parts of tagging. Two common approaches[2] for parts of speech tagging is

• Supervised pos tagging

• Unsupervised pos tagging

Supervised tagging technique requires a pre tagged corpora where as unsupervised

tagging technique do not require a pre tagged corpora. Both supervise and unsupervised

tagging can be of two types Rule based and stochastic. Rule based approaches use

contextual information to assign tags to unknown or ambiguous words. Rule based

system needs context rule for parts of speech tagging.

Stochastic tagging makes use of corpus. The most common stochastic tagging

technique uses a Hidden Markov Model (HMM). The probabilities are estimated from a

tagged training corpus or an untagged corpus in order to compute the most likely POS

tags for the word of an input sentence. Stochastic tagging technique may be of two types,

supervised stochastic technique and unsupervised stochastic technique. Supervised

stochastic tagging techniques use only tagged data. However the supervised method

requires large amount of tagged data so that high level of accuracy can be achieved.

Unsupervised stochastic technique not requires a pre tagged corpus but uses some

sophisticated computational methods.

**Brills tagger**

There are different approaches have been used for parts of speech tagging. Some

previous work has focused on rule based technique worked by Brill. Brills tagger[1]

most widely used parts of speech tagger used for English and also developed for some

other languages. The rule-based Brills part of speech tagger can be said to be a hybrid

approach, because it first uses statistical techniques to extract information from the

training corpus and then uses a program to automatically learn rules which reduces the

faults that would be introduced by statistical mistakes. Brills tagger is a combination of

both rule based and probabilistic model.

**Bangla Corpus**

In linguistics a **corpus** (plural *corpora*) or **text corpus** is a large and structured set of texts. Corpus is considered as basic resource for language analysis and research. This reflects both ideological and technological change in the area of language research. So for Bangla language analysis and research work it is necessary to have a complete Bangla linguistic corpus. In the recent year’s research work and linguistic analysis for Bangla language for different purposes is a most widely field.

**Existing work**

Parts of speech tagging for bangla language, in recent years are an interesting field

of research. Some works are going on in India and Bangladesh for parts of speech tagging

for bangla language. In India research team from IITKP is working on POS tagging for

bangla language. They developed a system which they called as a Hybrid tagging model.

First they outlined their training method which is based on partially supervised learning.

They run it with a number of iterations and process the untagged data by a supervised

training method from tagged data. In training session they used Baum-Welch algorithm.

They claim for 95% accuracy for tagging. In Bangladesh research team from Center for

Research on Bangla Language Processing (CRBLP) did some work on bangla parts of

speech tagging. Some work based on Maximum entropy markov model from Brac

University has been done, but the result is not quite satisfactory.

Bangla is one of the most widely spoken languages in the world with more than

200 million native speakers around the world. The grammar and morphological rules are

derived from sanskrit. In our proposed model parts of speech tagging is done using rule

based supervised approach

**Tag set**

Tag name Tag key word

|  |  |
| --- | --- |
| Noun | NN |
| Pronoun | PRP |
| Verb | VM |
| Adjective | ADJ |
| Adverb | ADV |
| Conjunction | CC |
| Preposition | PPS |
| Interjection | INT |

Table: Tag Set used for Tagging

**Stemming :** Stemming is a process by which we can reduce a part of a word to get the stem, base or root. Stemming process for a word takes us to origin of a word.

**Example-** Stemming/stemmed/stemmer -> Stem

Working/working/works -> Work

Here for the root word stem and work has several form. But through this stemmer we get the desire root or base word. Stemming is used in spelling checker and information retrieval systems to improve performance. For spellings checkers and search engine specifically.

**Several stemming algorithm**

There are several types of stemming algorithms which differ in respect to performance and accuracy and how certain stemming obstacles are overcome.

* **Brute-force algorithms:**

Brute force stemmers employ a lookup table which contains relations between root forms and inflected forms The algorithm is only accurate to the extent that the inflected form already exists in the table.

* **Suffix-stripping algorithms:**

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, typically smaller lists of “rules” are stored which provide a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

* if the word ends in 'ed', remove the 'ed'
* if the word ends in 'ing', remove the 'ing'
* if the word ends in 'ly', remove the 'ly'
* **Lemmatisation algorithms:** A more complex approach to the problem of determining a stem of a word is lemmatisation. This process involves first determining the part of speech of a word, and applying different normalization rules for each part of speech. The part of speech is first detected prior to attempting to find the root since for some languages, the stemming rules change depending on a word's part of speech.
* **Stochastic algorithms:**

Stochastic algorithms involve using probability to identify the root form of a word. Stochastic algorithms are trained (they "learn") on a table of root form to inflected form relations to develop a probabilistic model. This model is typically expressed in the form of complex linguistic rules, similar in nature to those in suffix stripping or lemmatisation. Stemming is performed by inputting an inflected form to the trained model and having the model produce the root form according to its internal rule set.

* **Hybrid approaches:**

Hybrid approaches use two or more of the approaches described above in unison. A simple example is a suffix tree algorithm which first consults a lookup table using brute force. However, instead of trying to store the entire set of relations between words in a given language, the lookup table is kept small and is only used to store a minute amount of "frequent exceptions" like "ran -> run". If the word is not in the exception list, apply suffix stripping or lemmatization and output the result.

* **Affix stemmers:**

In linguistics, the term affix refers to either a prefix or a suffix. In addition to dealing with suffixes, several approaches also attempt to remove common prefixes. For example, given the word *indefinitely*, identify that the leading "in" is a prefix that can be removed. Many of the same approaches mentioned earlier apply, but go by the name affix stripping.

**Matching algorithms**

Matching Algorithms: Such algorithms use a stem database (for example a set of documents that contain stem words). These stems are not necessarily valid words themselves (but rather common sub-strings). In order to stem a word the algorithm tries to match it with stems from the database, applying various constraints, such as on the relative length of the candidate stem within the word.

|  |
| --- |
| stemmer_alg.png |

**Fig: Stemming algorithms**

**Bengali stemmer**

In the field of English stemmer most successful work has been done. But in the field of Bengali still it is in the research level. In Bengali language a word can be stemmed by removing affixes (suffixes or prefixes) producing a root form called a stem that often closely approximates the root morpheme of a word. Usually those modifiers are Bivokti, Kal, Protyoi, Uposhorgo and Sondhi. Here the Uposhorgo is the prefix and rest of modifiers is suffixes.

# Prefix

A **prefix** is an [affix](http://en.wikipedia.org/wiki/Affix) which is placed before the [stem](http://en.wikipedia.org/wiki/Stem_%28linguistics%29) of a word. Particularly in the study of [Semitic languages](http://en.wikipedia.org/wiki/Semitic_languages), a prefix is also called a **preformative**, because it alters the form of the words to which it is affixed.

# Suffix

In [linguistics](http://en.wikipedia.org/wiki/Linguistics), a **suffix** (also sometimes called a *postfix* or *ending*) is an [affix](http://en.wikipedia.org/wiki/Affix) which is placed after the [stem](http://en.wikipedia.org/wiki/Stem_%28linguistics%29) of a word. Common examples are case endings, which indicate the [grammatical case](http://en.wikipedia.org/wiki/Grammatical_case) of nouns or adjectives, and verb endings, which form the [conjugation](http://en.wikipedia.org/wiki/Grammatical_conjugation) of verbs.

**How Bangla words build:**

We have analyzed morphological rule of bangla words. There are several methods of building bangla words:

**বিভক্তি**

শিশুকে=(শিশু+কে বিভক্তি)

ছেলেরা=(ছেলে+রা বিভক্তি

পানিতে=(পানি+তে বিভক্তি)

**কাল**

করছিলেন=(কর+ছিলেন)

করিয়েছি=(কর+ ইয়েছি)

করলাম=(কর+লাম)

**নির্দেশক**

টাকাটা=(টাকা+টা)

বাড়িটি=(বাড়ি+টি)

বইখানি=(বই+খানি)

**প্রত্যয়**

খেলনা=(খেল+অনা)

বাঘা=(বাঘ+আ)

দুধওয়ালা=(দুধ+ওয়ালা)

**উপসর্গ**

অজপুকুর=(অজ+পুকুর)

পাতিহাঁস=(পাতি+হাঁস)

আবছায়া=(আব+ছায়া)

**Some examples of Suffixes which change a Bangla word POS:**

We have made a list of suffixes which change a Bangla word POS. Here are given some examples:

চড়্(verb)+ আই=চরাই(noun)

ডুব্(verb)+ অন্ত= ডুবন্ত(adj)

ধোঁকা(noun)+ বাজ= ধোঁকাবাজ(adj)

**Some examples of Suffixes which do not change a Bangla word POS:**

We have made a list of suffixes which do not change a Bangla word POS. Here are given some

exampels

গরু(noun)+গুলি=গরুগুলি(noun)

কর(verb)+বেন=করবেন(verb)

পাতি+শিয়াল(noun)=পাতিশিয়াল(noun)

**Rules:**

We have made some rules for stemming which is stored in a rule file. Here are given some examples of rules:

**ে bivokti unchanged isnoun noun**

**য় bivokti unchanged isnoun noun**

**তে bivokti unchanged isnoun noun**

**া lingo unchanged isadj adj**

**ী lingo unchanged isadj adj**

**ি kal unchanged isverb verb**

**ে kal unchanged isverb verb**

**েছিলাম kal unchanged isverb verb**

**িয়া prottoy changed isverb adj**

**িয়ে prottoy changed isverb adj**

**িষ্ণু prottoy changed isverb adj**

**ি prottoy changed isadj adj**

In “ ে bivokti unchanged isnoun noun”, here ে is a bivokti, using bivokti the parts of speech of the stemmed word may not changed. If the stemmed word is noun, the word connected with bivokti is also noun.

In “িষ্ণু prottoy changed isverb adj” here িষ্ণু is a prottoy. Using prottoy with the stemmed word the parts of speech may change. If stemmed word is verb, the word connected with prottoy is an adjective.

**Procedure of our stemming**: We have implemented a stemmer which is a mixture procedure of “Affix stemmer and matching algorithm”. Because we used a Bengali corpus dictionary and modifiers (are Bivokti, Kal, Protyoi, Uposhorgo) from files.

Stem word in corpus

Found in

Corpus?

Stem word in corpus

Tag from corpus

Check rules file & tag

Modifier change stem word

Tag as unknown

Tag from corpus

Tag as unknown

Stem success?

Tag from corpus

Try to stem

no yes

no yes

Modifier

Prefix/ Suffix?

Prefix suffix

Yes no yes no

Yes no

Check rules and tag

Fig: Flow chart of stemming procedure

**Pseudo code for stemming procedure:**

If(input word found in corpus)

{

Tag from corpus;

}

Else

{

Stem();

If(stem() return false)

{

Tag unknown;

}

Else

{

If(modifier==prefix)

{

If(stem word found in corpus)

{

Tag from corpus

}

Else

{

Tag unknown;

}

Else

{

If(stem word found in corpus)

{

If(modifier change stem word)

{

Check rule file and tag;

}

Else

{

Tag from corpus;

}

}

Else

{

Check rules and tag;

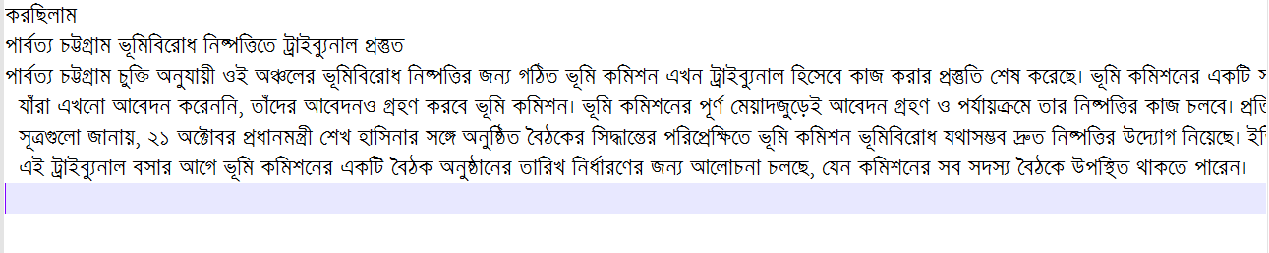
}

}

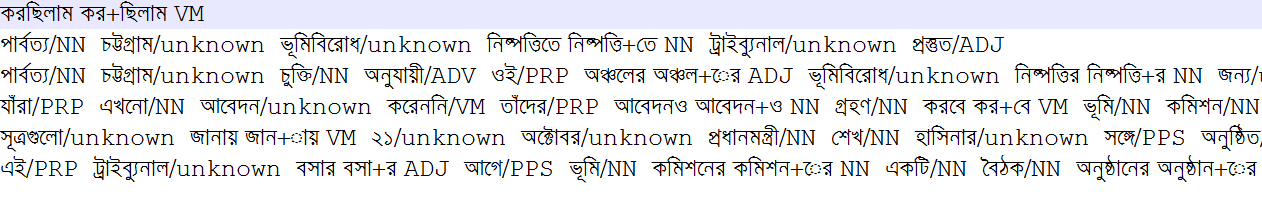
}

}

**Input:**

****

**Output:**

****

**Proposal:**

**Some Syntax Rules of Bangla Sentence:**

1. **বিশেষণ পদ বিশেষ্যর আগে বসে। যথাঃ সুন্দর ছবি, লাল জামা।**
2. **সর্বনামের বিশেষণ সাধারানত পরে বসে। যথাঃ তিনি বড় লোক।**
3. **ক্রিয়া বিশেষণ সাধারনত ক্রিয়ার আগে বসে। যথাঃ সে জোরে হাটে।**
4. **সমাপিকা ক্রিয়া সাধারানত বাক্যের শেষে বসে। যথাঃ আমি রোজ সকালে হাটি।**
5. **বাংলা বাক্যে কর্তা, কর্ম, ক্রিয়া সাধারনত এই ক্রমানুসারে হয়ে থাকে। যথাঃ আমি বই পড়ি।**

**Exceptional:**

**স্বাভাবিকঃ তুমি রোজ বিকালে আমাকে ফুল দিয়ে যাও**

**সম্মুখনঃ রোজ বিকালে তুমি আমাকে ফুল দিয়ে যাও।**

**সম্মুখনঃ আমাকে রোজ বিকালে তুমি ফুল দিয়ে যাও।**

**স্বাভাবিকঃ খুব কম ছেলেই ঘটকের চোখে পাত্রী দেখে বিয়ে করতে চায়।**

**সম্মুখনঃ ঘটকের চোখে পাত্রী দেখে খুব কম ছেলেই বিয়ে করতে চায়।**

**স্বাভাবিকঃ মনোজগতে বাইরের সৌন্দর্য এসে পৌঁছল।**

**পদ বিপর্যয়ঃ বাইরের সৌন্দর্য এসে পৌঁছল মনোজগতে।**

**#বাংলা ভাষায় একই বিশেষ্য ও বিশেষণ রূপে ব্যবহৄত হতে পারে।**

**যথাঃ ভাল বাড়ি পাওয়া কঠিন। (ভাল এখানে বিশেষণ)**

**আপন ভাল সবাই চায়। (ভাল এখানে বিশেষ্য)**

**মন্দ কথা বলতে নেই। (মন্দ এখানে বিশেষণ)**

**এখানে কি মন্দটা তুমি দেখলে? (মন্দ এখানে বিশেষ্য)**

**Solution:**

**Some example of Bangla Sentences’ Chain:**

NN+CC+NN+CC+CC+CC+NN+VM+NN+

NNL+NN+NNL+CC+NNL+NN+

ADJ+NN+NN+CC+NN+ADJ+VM+ADV+NN+ADV+NN+ADJ+VM+VM+ADJ+NN+PRP+NN

For tagging unknown words we are proposing that in a chain file all chain rules will have to store. Then the system will have to match the input sentence with the Chain file. The unknown word which is most familiar with the chain rule will have to tag.

**Conclusion**

We proposed two approaches (stemming and chain rule) for parts of speech tagging .Our stemming procedure is a rule-based approach which can efficiently stem and tag bangla words.For tagging unknown words we proposed chain rule for bangla POS tagging.The POS tagging is useful for bangla search engine, spell checking and other natural language processing applications.

**References**

1.” A simple Rule based parts of speech tagger” by Eric Brill

2. ”A hybrid model for Parts of speech tagging and its application in Bengali” by

Sandipan dandapat, Sudeshna sarkar and anupam basu.

3. “Syntactic parts of speech tagging guideline for bangla text” by Altaf Mahmud and

Mumit khan.

4. “Brown corpus” by henry Kucera and nelson francis.

5.Barnan Das & Tanmoy Pal, Development of Bengali Language Stemmer 2009,

Indian Institute of Technology, Kharagpur.

6.Morphological Stemming Cluster Identification for Bangla, Jadavpur University.

7.Porter Stemmer Algorithm,1980.

8.বাংলা ভাষার ব্যকরণ-মুনীর চৌধুরী,মোফাজ্জল হায়দার চৌধুরী

9.বাক্যতত্ত্ব-হুমায়ুন আজাদ

10.”Research Report on Bangla Verb and Noun Morphological analysis” by Md.Zahurul Islam, Center for Research on Bangla Language Processing (CRBLP),BRAC University.