Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 6; October-December, 2016, pp. 517-521

http://www.krishisanskriti.org/Publication.html

Parsing and Part-of-Speech tagging for Assamese Texts

Bipul Roy[1] and Bipul Syam Purkayastha[2]

[1]Scientist-B, NIELIT Guwahati Kokrajhar Ext. Centre and Research Scholar, Department of Computer Science, Assam University
[2]Department of Computer Science, Assam University
E-mail: [1]bipul.roy@nielit.gov.in, [2]bipul_sh@hotmail.com

Abstract—*Assamese is one of the morphologically rich and highly inflectional languages of North-East India. But till date a very less number of works have been done in terms of Natural Language Processing (NLP) and digitization of Assamese language corpora. The two main factors that determine the syntactic category of an Assamese word are the lexical information directly related to the category of an Assamese word and the contextual information related to the environment in which the word is used. Parsing and Part-of-Speech tagging of any Assamese word can provide necessary information to understand the syntax and semantics of the Assamese grammar. After studying the various standard tagset developed by different language researchers, we have developed an Assamese tagset improvising the BIS POS tagset with 31 tags which contain 11 top level categories and their respective subtypes. Here we have tried to tag and parse the Assamese texts using Context Free Grammar (CFG) in Python with Natural Language Toolkit (NLTK). This is just a simple model where we have used lesser numbers of tags for parsing to minimize the errors and for better productivity. This model parser can parse almost all the simple sentences of Assamese texts with an accuracy of 98.6%, though in case of complex sentences it faced difficulty to parse. In this research work an attempt is being made to develop Assamese tagset and Assamese tagged corpus for the better understanding of the Assamese language structure and its grammar.*

Keywords*: NLP, CFG, NLTK, Parsing, Assamese language, Python*

1. INTRODUCTION

For the development and enrichment of all languages part of speech tagging plays a vital role. Part of speech tagging specially for the regional Indian languages can give an international and worldwide approach. On the basis of linguistic features for assigning appropriate tags for each word from an input corpus is called part-of-speech tagging, while parsing is a term used to describe the process of automatically building syntactic analysis of a sentence in terms of a given grammar. Among the different challenges of Part-of-Speech (POS) tagging and Parsing, the most difficult challenge is to resolve the ambiguity of words which occur because of using the same word in different contexts [11, 12].

1.1. Natural Language Processing & POS Tagging

A computer program that understands the human language as it is spoken is termed as Natural Language processing, whereas natural language is the language spoken by human. As human language is not specific always, so it is not a simple task to design an NLP application [3].

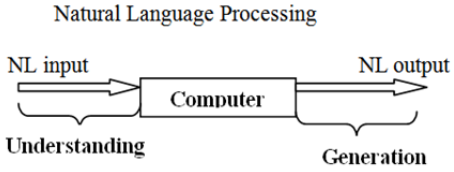

Fig. 1.1: Natural Language Processing (NLP)

Here the purpose of the system is to develop a program that tags correctly each and every word of the corpus. In terms of language automation, Part of Speech (POS) tagging is the process of automatic assignment of lexical categories or word class markers (Noun, Verb, Adjective, Adverb etc.) for each word in a sentence of a natural language. POS tagging, which is also called word-category disambiguation or grammatical tagging, is based on both the definition and the context (i.e. the relationship of the word with adjacent and related words in a phrase, sentence or paragraph). The input of the tagging process is the sequence of words of a natural language sentence and specified tag sets (a finite list of part of speech tags). The output is a single best part of speech tag category for each word in the sentence. We have designed Assamese tagset which contains 31 tagset incorporating Bureau of Indian Standard (BIS) tagset used for different Indian languages. We have also manually tagged around 10,000 words of Assamese texts using our designed tagset. The overall process of part of speech tagging for Assamese sentence is shown in the Fig. below.

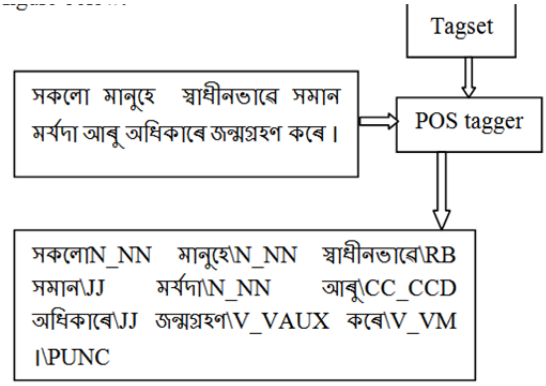

Fig. 1.2: the process of part of speech tagging of Assamese Text

In NLP different approaches have already been tried out to automate the process of POS tagging of English, European and a few South Asian languages. The automation process is carried out by either a large set of linguistic rules or by annotated corpus with comprehensive size. But such rules and corpora have been developed and available for English and European languages. However, if we look at the same scenario for Indian languages, we find out that not much work has been done. The main reason behind this is the unavailability of concise linguistic rules and large annotated corpus. For Assamese language, it still lacks significant research efforts in the area of NLP.

1.2. Challenges in POS Tagging

Natural languages are ambiguous by nature. The main problem in part of speech tagging is part of speech ambiguity. There may be many words which can have more than one tag. For example consider the **noun-verb** ambiguity of the word ***book*** in the sentences "*This is a* ***book***". Here the word ***"book"*** is noun and "I ask him to ***book*** *a ticket for me*". Here the word ***"book"*** is verb. To solve this problem one can consider the context instead of taking single word in the sentence. So the most challenging problems in POS tagging disambiguation are to determine the proper context and adequate features. Also for Indian languages which are highly inflectional and morphologically rich, automatic tagging is a challenging task. The occurrence of unknown words, (i.e. words that do not exist in the corpus) is also another issue which may create problem in POS tagging [11, 12].

Thus it is necessary to consider all the issues to develop an automatic part of speech tagger.

2. ASSAMESE LANGUAGE

Assamese is the part of the eastern group of the Indo-Aryan languages. Along with other eastern Indo-Aryan languages, Assamese evolved at least before 7th century A.D. from the Magadhi Prakrit. The current form of the Assamese script has been continuous development from the 5th century. Like other eastern Indo-Aryan languages, Assamese is also morphologically rich language. There are mainly two dialects in Assamese viz. Kamrupi and Goalpariya [4, 5].

For most languages that have a major class of nouns, it is possible to define a basic word order in terms of subject(S), verb (V) and object (O). There are six theoretically possible basic word orders: SVO, SOV, VSO, VOS, OVS, and OSV. Of these six, however, only the first three normally occur as dominant orders. If constituents of a sentence can occur in any order without affecting the gross meaning of the sentences (the emphasis may be affected), such types of languages are known as free word order languages. Warlpiri, Russian and Tamil are examples of free word order languages.

Typical Assamese sentences can be divided into two parts: Subject(S) and Predicate (P).

Though, there is a pressing necessity to develop an automatic Part of Speech tagger for Assamese language, our approach of Assamese tagger is just an initial step towards the long journey of NLP. Our tagger can tag the following types of Assamese corpus successfully providing the output as:

সকলো\N_NN\JJ মানুহে\N_NN স্বাধীনভাৱে\RB সমান\JJ মর্যদা\N_NN আৰু\CC_CCD অধিকাৰে\JJ জন্মগ্রহণ\V_VAUX কৰে\V_VM ।\PUNC সিহঁতৰ\PR_PRP বিবেক\N_NN আৰু\CC_CCD বুদ্ধি\N_NN আছে\V_VM আৰু\CC_CCD সিহঁতে\ PR_PRP পৰস্পৰ\JJ\RB ভাতৃত্বৰে\JJ আচৰণ\N_NN কৰিব\V_VM লাগে\V_VM ।

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience. Therefore, they should act towards one another in a spirit of brotherhood.

POS tagging can be both manual as well as automatic. Though manual tagging is more accurate but it is time-consuming, long and continuous process. Hence, the automatic tagger is essential to speed up the process of POS tagging with less chance of errors and inconsistency. Various automatic POS taggers have been developed worldwide using linguistic rules, stochastic models and hybrid models. Automatic tagging is a challenging task for Indian languages which are highly inflectional and morphologically rich.

2.1 Assamese Tagset

A tag set consists of tags that are used to represent the grammatical information of the language. The number of tags that we use for a language depends upon the information that we want to represent using a tag. The granularity of a tagset depends on the requirement of researcher. In our designed tagset for Assamese language, there are 31 tags which contain eleven (11) top level categories and their respective subtypes [8, 9].

Table 1: Assamese version of the BIS Tagset

Sl	Category		Label	Annotation Convention	
	Top Level	Subtype Level			Assamese
1	Noun		N		বিশেষ্য
1.1		Common	NN	N_NN	জাতিবাচক
1.2		Proper	NNP	N_NNP	ব্যক্তিবাচক
1.3		Nloc	NST	N_NST	স্থানবাচক
2	Pronoun		PR		সর্বনাম
2.1		Personal	PRP	P_PRP	ব্যক্তিবাচক
2.2		Reflexive	PRF	P_PRF	আত্মবাচক
2.3		Relative	PRL	P_PRL	সম্বন্ধবাচক
2.4		Reciprocal	PRC	P_PRC	পাৰস্পৰিক
2.5		Wh-word	PRQ	P_PRQ	প্রশ্নবোধক সর্বনাম
3	Demonstrative		DM		নির্দেশবোধক
3.1		Deictic	DMD	DM_DMD	প্রত্যক্ষ নির্দেশক
3.2		Relative	DMR	DM_DMR	সম্বন্ধবাচক
3.3		Wh-word	DMQ	DM_DMQ	প্রশ্নবোধক অব্যয়
4	Verb		V		ক্রিয়া
4.1		Main	VM	V_VM	মুখ্য ক্রিয়া
4.2		Auxiliary	VAUX	V_VAUX	সহায়কাৰী ক্রিয়া
5	Adjective		JJ		বিশেষণ
6	Adverb		RB		ক্রিয়া বিশেষণ
7	Postposition		PSP		অনুসর্গ
8	Conjunction		CC		সংযোজক
8.1		Co-ordinator	CCD	CC_CCD	সমন্বয়ক
8.2		Subordinator	CCS	CC_CCS	
9	Particles		RP		আনুষংগিক অব্যয়
9.1		Default	RPD	RP_RPD	
9.2		Classifier	CL	RP_CL	নির্দিষ্টবাচক
9.3		Interjection	INJ	RP_INJ	বিস্ময়বোধক
9.4		Intensifier	INTF	RP_INTF	
9.5		Negation	NEG	RP_NEG	নঞর্থক
10	Quantifiers		QT		পৰিমাণবাচক
10.1		General	QTF	QT_QTF	সাধাৰণ
10.2		Cardinals	QTC	QT_QTC	সংখ্যাবাচক
10.3		Ordinals	QTO	QT_QTO	ক্রমবাচক সংখ্যাবাচক শব্দ
11	Residuals		RD		
11.1		Foreign Word	RDF	RD_RDF	বিদেশী শব্দ
11.2		Symbol	SYM	RD_SYM	প্রতীক
11.3		Punctuation	PUNC	RD_PUNC	যতি চিন
11.4		Unknown	UNK	RD_UNK	অজ্ঞাত
11.5		Echowords	ECH	RD_ECH	ধ্বন্যাত্মক শব্দ

3. DEFINITION OF PARSING

Parsing is a term used to describe the process of automatically building syntactic analysis of a sentence in terms of a given grammar. So parsing analyzes the input sentence and determines its grammatical structure. A parser is a computer program that carries out this task [2]. The output of a parsing is logically equivalent to a tree. For a sentence if a parser successfully generates a parse tree and the leaf nodes of the tree are tokens of input sentence, then the sentence is grammatically correct [1, 2, 6, 7, 10, 13].

3.1. Proposed System Architecture

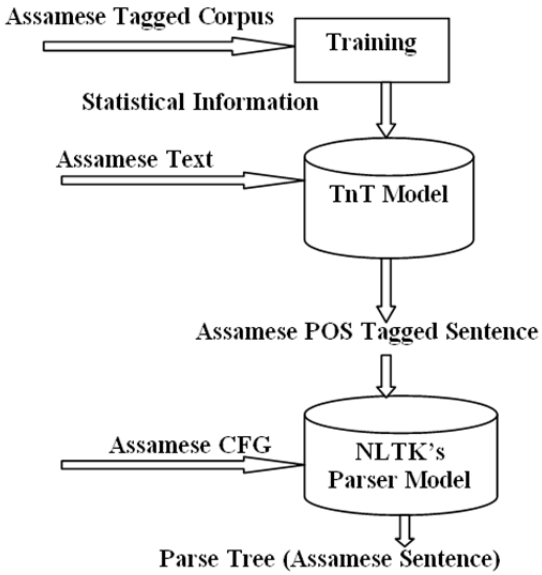

Fig. 3.1 System Architecture

In this system we have used Assamese tagged corpus, the training dataset (manually tagged) as input to the TnT tagger to create Assamese model. Then we have supplied the test dataset Assamese plain text as input to the model to get the tagged Assamese corpus.

We have implemented & designed Assamese parser using Context free Grammar of Assamese language in Python with Natural Language Toolkit to generate parse tree for the input sentences one by one.

Assamese parser is a tool which takes Assamese sentence and verifies whether the given Assamese sentence is correct or not according to Assamese language grammar. Parsing is important for Natural Language Processing tools. Assamese parser uses the Recursive-descent parsing algorithm for Parsing of Assamese language. It parses whole sentence and generates a parse tree. Recursive-descent parsing is a kind of top-down parser built from a set of mutually recursive procedures (or a non-recursive equivalent) where each such procedure usually implements one of the productions of the grammar. Thus the structure of the resulting program closely mirrors that of the grammar it recognizes.

We are going to discuss the following important terms used in our designed model.

3.2. TnT Tagger

Trigram n Tags (TnT) is an efficient statistical POS tagger, which is completely based on Hidden Markov Model (HMM) and used some optimizing techniques for smoothing and handling unknown words [10].

3.3. Context free grammars

The idea of a context-free grammar (CFG) should be familiar from formal language theory. A CFG has four components, described here as they apply to grammars of natural languages:

1. a set of non-terminal symbols (e.g., S, VP), conventionally written in uppercase;

2. a set of terminal symbols (i.e., the words), conventionally written in lowercase;

3. a set of rules (productions), where the left hand side (the mother) is a single non-terminal and the right hand side is a sequence of one or more non-terminal or terminal symbols (the daughters);

4. a start symbol, conventionally S, which is a member of the set of non-terminal symbols [10].

3.4. Python

The Python programming language is a dynamically-typed, object-oriented interpreted language. Although, its primary strength lies in the ease with which it allows a programmer to rapidly prototype a project, its powerful and mature set of standard libraries make it a great fit for large-scale production-level software engineering projects as well. Python has a very shallow learning curve and an excellent online learning resource [10].

3.5. Natural Language Toolkit

Although Python already has most of the functionality needed to perform simple NLP tasks, it's still not powerful enough for most standard NLP tasks. This is where the **Natural Language Toolkit (NLTK)** comes in [10].

NLTK is a collection of modules and corpora, released under an open source license that allows students to learn and conduct research in NLP.

The most important advantage of using NLTK is that it is entirely self contained.

Not only does it provide convenient functions and wrappers that can be used as building blocks for common NLP tasks, it also provides raw and pre-processed versions of standard corpora used in NLP literature and courses.

Some sentences having both noun phrase and verb phrase exist. For example "মই ভাত খাওঁ"

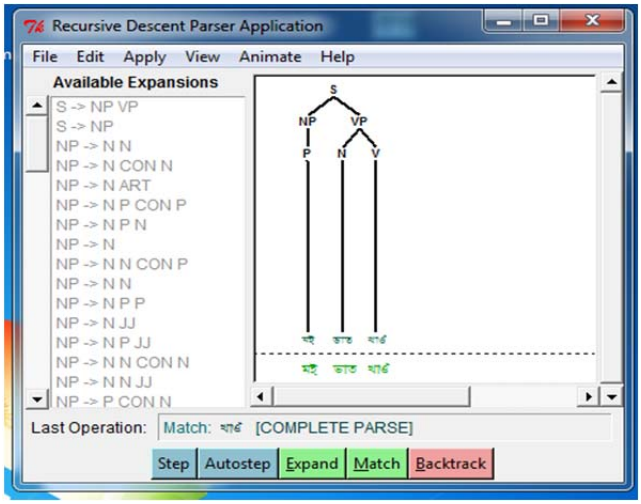

Fig. 3.2. Parse tree for "মই ভাত খাওঁ"

Some other sentences having only noun phrase exist. For example "ধুনীয়া এজনী ধুনীয়া ছোৱালী". But only verb phrase cannot create a complete sentence in Assamese language. Otherwise there are grammatical errors in the sentence

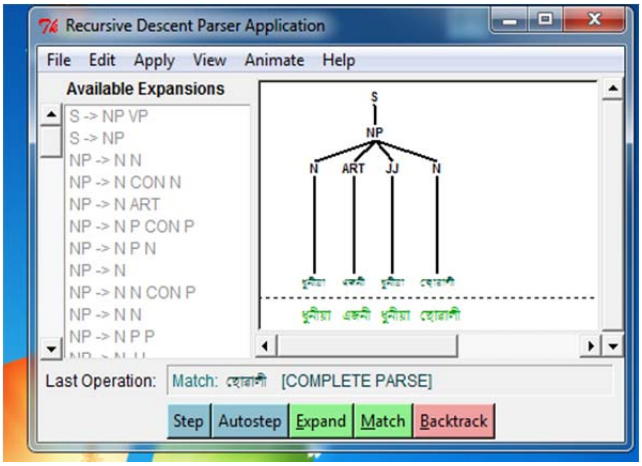

Fig. 3.3. Parse tree for "ধুনীয়া এজনী ধুনীয়া ছোৱালী"

4. OUR CONTRIBUTIONS

1. Studies and analysis of Linguistic features of Assamese language.
2. Development of tagset for Assamese by taking into consideration of both language features and the morpho-syntactic features.
3. Development of annotated corpus.
4. Development of POS tagger for Assamese language.

5. Development of Recursive Descent Parser for Assamese language.

We have implemented a benchmarked model to understand the aspects and challenges of POS tagging and Parsing of Assamese language. In this model the tag probabilities depend only on the current word and using Recursive Descent Parser to parse the input Assamese sentences with the help of Context free grammar.

5. CONCLUSIONS & FUTURE WORKS

In the field of NLP, parsing and tagging of Assamese text has a great impact. Such kind of NLP works can help many students, teachers, linguistics researchers and any person willing to learn about Assamese language and its grammar. Still there is huge scope awaiting to be explored in this field for the betterment and flourishment of the Assamese language.

Here in this work we have used 31 tags for tagging but for parsing of Assamese sentences, we have used less number of tags. So we are trying to include the remaining tags for parsing in our future works. Also we will try to deal with the ambiguities of Assamese texts in future.

REFERENCES

[1] http://en.wikipedia.org/wiki/parse tree.

[2] Alfred V. Aho and Jeffrey D. Ullman. Principles of Compiler Design. Narosa publishing House, 1995.

[3] James Allen. Natural Language Understanding. Pearson Education, Singapur, second edition, 2004.

[4] Hem Chandra Baruah. Assamiya Vyakaran. Hemkosh Prakashan, Guwahati, 2003.

[5] D. Deka and B. Kalita. Adhunik Rasana Bisitra. Assam Book Dipot, Guwahati, 7th edition, 2007.

[6] T.V. Geetha K. Saravanan, Ranjani Parthasarathi. Syntactic Parser for Tamil. Resource Center for Indian Language Technology Solutions-Tamil, School of Computer Science and Engineering, Anna University, Chennai, 2003.

[7] Stephen G. Pulman. Basic Parsing Techniques: an introductory survey. Pergamon Press and Aberdeen University Press, University of Cambirdge Computer laboratory, 1991.

[8] Bipul Roy, Bipul Syam Purkayastha: Annotating Assamese Corpus using the Standard POS Tagset, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016

[9] S. Utpal, J. Kalita, and R. Das. Unsupervised Learning of Morphology for Building a Lexicon for Highly Inflectional Language. 2002.

[10] B.M. Sagar, G. Shobha and P. Ramakanth Kumar: Solving the Noun Phrase and Verb Phrase Agreement in Kannada Sentences, International Journal of Computer Theory and Engineering, Vol. 1,No. 3, August, 2009 1793-8201

[11] N. Saharia, U. Sharma, J. Kalita. A First Step towards Parsing of Assamese Text. Language in India www.languageinindia.com1:5 May, 2009.

[12] P. Arunmozhi, L. Sobha, B. Kumara Shanmugam. Part of Speech Tagger for Tamil. AU-KBC Research Centre, MIT campus of Anna University, Chennai - 44.

[13] P.J. Antony, K.P. Soman. Computational Morphology and Natural Language Parsing for Indian Languages: A Literature Survey**.** International Journal of Scientific & Engineering Research Volume 3, Issue 3, March-2012 1 ISSN 2229-5518