Annotating Assamese Corpus using the Standard POS Tagset

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

Scientist B, NIELIT, Guwahati Kokrajhar Ext. Centre, Assam, India [1]

Professor, Department of Computer Science, Assam University, Assam, India [2]

Abstract: Assamese is the official language of the Indian state of Assam and is about 25 million native speakers. But, being a regional language, it is highly lacking in language resources like corpus, language technology tools, guidelines etc till date. As the digitization of Assamese corpus, after it was tagged at the Part-of-Speech (POS) level, can help tremendous in the fields of various Natural Language Processing (NLP) applications, linguistic studies, various linguistic research works, etc. So, the development of annotated Assamese corpus has become unavoidable task now-a-days.

Keywords: Assamese, POS, BIS, NLP.

1. INTRODUCTION

For any type of NLP tasks as Linguistic researchers, corpus plays a very crucial role. Digital corpus is part and parcel for NLP tasks like POS tagging, machine learning, theoretical linguistics, etc, compared to other major Indian Languages, Assamese has negligible amount of digital corpus so far. Such lack in digital corpus of Assamese language hampers the research works to a great extent. Another hurdle in terms of POS tagging of Assamese is the flexibility of the word of Assamese language. This makes the tagging task more complex.

2. LINGUISTIC CHARACTERISTICS OF ASSAMESE

Just like other indo-Aryan languages Viz Bengali, Hindi etc, Assamese is also relatively free word order though the predominant word order is SOV (Subject-Object-Verb). In Assamese language with respect to the numbers and genders agreement, verb and its subject can not affect their respective forms [5]. For example

(i) ল’ৰাজন পথাৰত খেলি আছে ।

(ii) ল’ৰাটো পথাৰত খেলি আছে । In singular form of sentences

(i) ল’ৰাবোৰ পথাৰত খেলি আছে ।

(ii) ল’ৰাবিলাক পথাৰত খেলি আছে ।

(iii) ল’ৰাহঁতে পথাৰত খেলি আছে । in plural form of sentences

And in case of gender also, the change on verb and its subject occur, for example:

(i) ল’ৰাজন vs ছোৱালজনী

The suffix “জনী” is only used in term of females.

(ii) The pronoun “সি” is used to indicate males.

A. BIS tagset

There are a number of standard tagset IIIT (ILMT) tagset, BIS tagset, LDC-IL tagset, AUKBC tagset, Tamil tagset, JNU-Sanskrit tagset (JPOS), Sanskrit consortium tagset (CPOS), MSRISanskrit tagset (IL-POSTS) and CIIL Mysore tagset available now-a-days developed by various language researchers and BIS POS tagset is one of them [1, 2, 3, 6, 7, 8] .

The BIS POS tagset is designed under the banner of Bureau of India standards by the Indian languages corpora initiatives (ILCI) group to meet the challenges of NLP. A proper tagset is one of the important pillars of NLP. Moreover these standard tagsets are bit flexible in nature and language researchers can modify the tagsets as per her/his requirements in their research work. In this particular tagset for Assamese language, there are 31 tags which contain eleven top level categories and their respective subtypes.

3. BIS POS TAGSET FOR ASSAMESE

The BIS tagset was used for tagging the Assamese ILCI corpus consisting of a total of 577918 tokens. The written corpus in Assamese language will provide data from different domains such as literature, science, media, art etc and will help researchers to carry out their work.

The main objective of the ILCI project was to develop standard quality parallel annotated corpora for 11 Indian languages including English language to promote NLP research for Indian Languages [1, 4].

TABLE I: 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		সৰ্বনাম
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	প্ৰশ্নবোধক সৰ্বনাম
3	Demonstrative		DM		নিৰ্দেশবোধক
3.1		Deictic	DMD	DM_DMD	প্ৰত্যক্ষ নিৰ্দেশক
3.2		Relative	DMR	DM_DMR	সম্বন্ধবাচক
3.3		Wh-word	DMQ	DM_DMQ	প্ৰশ্নবোধক অব্যয়
4	Verb		V		ক্ৰিয়া
4.1		Main	VM	V_VM	মুখ্য ক্ৰিয়া
4.2		Auxiliary	VAUX	V_VAUX	সহাকাৰী ক্ৰিয়া
5	Adjective		JJ		বিশেষণ
6	Adverb		RB		ক্ৰিয়া বিশেষণ
7	Postposition		PSP		অনুসৰ্গ
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	নিৰ্দিষ্টবাচক
9.3		Interjection	INJ	RP_INJ	বিস্ময়বোধক
9.4		Intensifier	INTF	RP_INTF	
9.5		Negation	NEG	RP_NEG	নঞাৰ্থক
10	Quantifiers		QT		পৰিমাণবাচক
10.1		General	QTF	QT_QTF	সাধাৰণ
10.2		Cardinals	QTC	QT_QTC	সংখ্যাবাচক
10.3		Ordinals	QTO	QT_QTO	ক্ৰমবাচক সংখ্যাবাচক শব্দ
11	Residuals		RD		
11.1		Foreign Word	RDF	RD_RDF	বিদেশী শব্দ
11.2		Symbol	SYM	RD_SYM	প্ৰতীক
11.3		Punctuation	PUNC	RD_PUNC	যতি চিন
11.4		Unknown	UNK	RD_UNK	অজ্ঞাত
11.5		Ecowords	ECH	RD_ECH	ধ্বন্যাত্মক শব্দ

DOI 10.17148/IJARCCE.2016.5879

4. ANNOTATING OF THE ASSAMESE CORPUS

A standard tagset is one of the most essential parts of NLP research. Here, we are going to discuss briefly about the various features of Assamese and specific issues in tagging the Assamese text.

1. Noun (N): A noun is a word that functions as the name of some specific thing or set of things, such as living creatures, objects, places, actions, qualities, states of existence, or ideas.In the BIS scheme, noun has four (04) subcategories viz common, proper, verbal and noun location. But we have studies here (03) subcategories of Noun viz. common, proper, noun location which is best suited for Assamese language.

a. Common Noun (NN): Common nouns are words used to name general items rather than specific ones. For example, Lamp, chair, couch, TV, window, painting, pillow and candle – all of these items are named using common nouns.

মোৰ \PR দেউতা \NN এজন শিক্ষক \NN ।\PUNC

b. Proper Noun (NNP): Proper nouns have two distinct features: They name specific one-of-a-kind items, and they begin with capital letters, no matter where they occur within a sentence.

মই অহাকালি গুৱাহাটীত \NNP যাম ।

গুৱাহাটী \NNP এখন বিখ্যাত নগৰ ।

কালিলৈ মই গুৱাহাটীলৈ \NNP যাম ।

c. Noun Location (NST): Locative (abbreviated LOC) is a grammatical case which indicates a location.

মই যেতিয়া \NST স্কুললৈ গৈ আছিলোঁ, সি তেতিয়া \NST গা ধুই আছিল ।

2. Pronoun (PR): In linguistics and grammar, a pronoun is a word that substitutes for a noun or noun phrase. It is a particular case of a pro-form. There are five (05) subtypes of pronoun viz, Personal, reflexive, relative, reciprocal and wh-word.

a. Personal Pronoun (PRP): Personal pronouns are pronouns that are associated primarily with a particular grammatical person – first person (as I), second person (as you), or third person (as he, she, it). Personal pronouns may also take different forms depending on number (usually singular or plural), grammatical or natural gender, case, and formality. The term "personal" is used here purely to signify the grammatical sense; personal pronouns are not limited to people and can also refer to animals and objects.

মই \PRP আজি অফিচত নাযাওঁ ।

b. Reflexive Pronoun (PRF): Reflexive pronouns are used when a person or thing acts on itself, for example, John cut himself. In English they all end in -self or -selves and must refer to a noun phrase elsewhere in the same clause.

শ্ৰীকৃষ্ণ স্বয়ং \PRF অৰ্জুনৰ সাৰথি আছিল ।

মই স্বয়ং \PRF সভাত উপস্থিত আছোঁ ।

c. Reciprocal Pronoun (PRC): Reciprocal pronouns are a type of pronoun which can be used to refer to a noun phrase mentioned earlier in a sentence. The reciprocal pronouns found in English are one another and each other, and they form the category of anaphors along with reflexive pronouns (myself, yourselves, etc.).

তুমি যিজনক \PRC বিচাৰি ফুৰিছা, সি মোৰ বন্ধু হয় ।

d. Relative Pronoun (PRL): A relative pronoun marks a relative clause; it has the same referent in the main clause of a sentence that the relative modifies. An example is the English word which in the sentence "This is the house which Jack built." Here the relative pronoun which marks the relative clause "which Jack built", which modifies the noun house in the main sentence. Which has an anaphoric relationship to its antecedent "house" in the main clause?

যি \PRL **খাবলৈ** মন যায়, সেইটটোকে খোৱা ।

e. Wh-word (PRQ): An interrogative word or question word is a function word used to ask a question, such as what, when, where, who, whom, why, and how. They are sometimes called wh-words, because in English most of them start with wh- (compare Five Ws). They may be used in both direct questions (Where is he going?) and in indirect questions (I wonder where he is going). In English and various other languages the same forms are also used as relative pronouns in certain relative clauses (The country where he was born) and certain adverb clauses (I go where he goes).

Wh-pronoun like কেতিয়া, কিমান, কাহানি, কেনে etc fall in this category. Again, the words having suffix "বা" also fall in this category, e.g. কেতিয়াবা, কেনেবা, কিবা, etc

তুমি কি \PRQ খাবলৈ ভাল পোৱা মই জানো ।

3. Demonstrative (DM): The next top level category is of demonstrative. Demonstratives have the same form of the pronouns, but distributionally they are different from the pronouns as they are always followed by a noun, adjective or another pronoun. In this category only deictic, relative and wh-word subtypes fall.

a. Deictic (DMD): Deictic are mainly personal pronouns.

এইযোৰ \DMD কাপোৰ মোক দেউতাই দিছে ।

b. Relative (DMR): Relative demonstrative are non-distinguishable from relative pronouns, except for that a demonstrative is followed by a noun, pronoun or adjective. In DRL distance attribute is absent.

যি \DMR **মূলা** বাঢ়ে তাৰ দুপাততে চিন ।

c. Wh-word (DMQ): Wh demonstrative are non-distinguishable from wh pronouns, except for that a demonstrative is followed by a noun, pronoun or adjective. The change in the morphological form is not found.

কোনে \DMQ তোমাক ইয়ালৈ পঠিয়ালে?

4. Verb (V): The category of verb is somewhat complicated in this framework. It has main and auxiliary

divisions under subtype level 1 and finite, non-finite, infinite and gerund divisions under subtype level 2. Like Hindi Verbs, Assamese verbs are complicated. They often appear in a group of words containing the verb of predication the verb of predication along with light verbs and auxiliaries.

a. Main verb (VM): We apply the main verb tag for all the forms that express the main predication of the sentence. We do not use distinct tags for finite and non-finite as Assamese, like Hindi, does not have enough information at the word level. Following are some examples:

আমি এতিয়া ব্যাকৰণ পঢ়িছোঁ \VM ।

b. Auxiliary (VAUX): In Assamese, like in Hindi, the auxiliary verbs concatenate with either the verbal root or verbal inflected forms and they serve to signal distinctions of tense, aspect, mood and voice.

অফিচত যাবলৈ তেওঁ ঘৰৰ পৰা ওলাই আহিল \VAUX ।

5. Adjective (JJ): An adjective modifies a noun. Though adjectives are not always followed by nouns in Assamese, it can be used as a predicate too. The first kind is called an attributive adjective and the second type is called a Predicative adjective. An adjective can function as a noun if not followed by a modified noun; in that case it is called an absolute adjective. When they are used with their modified item, should be tagged as adjectives otherwise as nouns.

ব্ৰহ্মপুত্ৰ এখন ডাঙৰ \JJ নৈ ।

6. Adverb (RB): Only manner adverbs are to be tagged as Adverbs in this framework.

তেওঁ মানসিকভাৱেও \RB দুৰ্বল হৈ পৰিছিল ।

7. Postposition(PSP):
Case relations are expressed by postpositions in Assamese.

ৰামচন্দ্ৰ সীতাৰ সৈতে \PSP বনলৈ গৈছিল ।

8. Conjunction(CC):
Conjunction is a major category in the tagset and has coordinator, subordinator and quotative as subtypes. We have to first enlist the conjunctions in these subcategories and then tag accordingly.

a. Co-ordinator (CCD)
The conjunctions that join two or more items of equal syntactic importance will be assigned CCD label. The list mainly includes

মই আৰু \CCD সি একেলগে স্কুলত গৈছিলো ।

b. Subordinator (CCS): The conjunctions that introduce a dependent clause are subordinators.

তথাপি \CCS সি বাপেকৰ মৰমৰ পৰা বঞ্চিত নহ’ল ।

9. Particle (RP): Particles have many a role to play in the language. In the tagset, there are default, classifier, interjection, intensifier and negation subtypes of the particles category.

a. Default (RPD): Words that express emotion are interjections, and also the particles which we use for getting the attention of people.

আপুনি এতিয়া আহিব বুলি \RPD মই ভবাই নাছিলো ।

b. Classifier (CL): The classifiers are combined with all types of nouns and numerals occurring in the language resulting in the combinations of the following type of grammatical construction. Unlike Hindi, Assamese has a couple of classifiers.

মানুহজন \CL ঢুকালে ।

c. Interjection (INJ): Words that express emotion are interjections, and also the particles which we use for getting the attention of people.

ও ৰাম! \INJ এইটো তই কি কৰিলি!

d. Intensifier (INTF) :Adverbial elements with an intensifying role are intensifiers. They could be both, either positive or negative will fall in this category.

ক্ৰিকেট এটা অত্যন্ত \INTF জনপ্ৰিয় খেল ।

e. Negation (NEG): The indeclinables which are used for negative meaning are treated under this category.

মই আজি অফিচত যোৱা নাই \NEG ।

10. Quantifiers (QT): A quantifier is a word which quantifies the noun, i.e., it expresses the noun’s definite or indefinite number or amount. The Quantifier category includes general, cardinal, and ordinal subtypes. These terms are equally applicable to both types of quantifiers: written in words (like five, fifth etc.) and in digits (like 5, 5th etc.).

a. General quantifier (QTF): This tag is for general kind of quantifiers.

সি পঢ়াশোনাত বহুত \QTF পিচ পৰা আছিল ।

b. Cardinal quantifier (QTC): The numbers which quantify objects are cardinal quantifiers.

প্ৰত্যেক ছাত্ৰৰে অন্ততঃ দুটাকৈ \QTC **পোছাক** আছিল ।

c. Ordinal quantifier (QTO): Quantifiers that specify the order in which a particular object is placed in a given world are ordinal quantifiers.

এই প্ৰস্তাবটো প্ৰথমে \QTO **মই** দাঙি ধৰিছিলো ।

11. Residuals (RD): Residual as a major category in this tagset has five subtypes; foreign word, symbol, punctuation, unknown and echo words as subtypes.

a. Foreign (RDF): In this framework a word is considered a foreign one if it is written in a script other than Devanagari script.

সি স্কুলৰ \RDF পৰা আহি খেলিবলৈ গ’ল ।

b. Symbol (SYM): The symbol subtype is for symbols like $, %, # etc.

সি পৰীক্ষাত ৭৬%\SYM নম্বৰ লৈ উত্তীৰ্ণ হৈছে ।

c. Punctuation (PUNC): Only for punctuations like?, ; , “ , । , etc., so other symbols than punctuations will be tagged as Symbol.

ৰীণা প্ৰথম শ্ৰেণীৰ ছাত্ৰী ।\PUNC

d. Echo words (ECH): Echo words are two words that occur together and the second one has no meaning on its own and it cannot occur on its own. It enhances the meaning of the word with which it occurs.

চাইকেলটো আজি বেচিকে কেৰ-কেৰ \ECH শব্দ কৰি আছে।

5. COLLECTION OF ASSAMESE CORPUS

For a new language researcher in Assamese, it is a very difficult task to collect annotated corpus. Also there are only a few resources of unannotated Assamese corpus, for example Online Assamese newspapers.

The digitization of Assamese language can be boasted tremendously if the young generations can be encouraged to write blogs in their mother tongue and also if the local writers of the language can be provided with some incentive like establishing some rewards and honours to write in their language.

As the Assamese literature is very rich and there are so many renowned writers in Assamese, it can be said that we can only lack in manpower in terms of digitization of Assamese corpus. This issue can be resolved to an extent if the Assam state govt. can fund initiatives of developing Assamese corpus.

6. CONCLUSION

Here in this paper, we have discussed the POS tagging of Assamese language and its related issues like collection of annotated corpus, grammatical difficulties, faced during POS tagging in machine, etc. So, far, we have found that getting the annotated Assamese Corpus is the toughest challenge faced by the language researchers and a few schemes to resolve this issue. Though the BIS tagset is not flawless, the language experts can minutely examine the tagset and can always suggest necessary modification for the refinement of this tagset.

REFERENCES

[1] Madhav Gopal: Annotating Bundeli Corpus using the BIS POS Tagset, Centre for Linguistics, SLL & CS, Jawaharlal Nehru University, New Delhi.

[2] Nitish Chandra, Sudhakar Kumawa, Vinayak Srivastava: VARIOUS TAGSETS FOR INDIAN LANGUAGES AND THEIR PERFORMANCE IN PART OF SPEECH TAGGING, Department of Computer Science & Engineering, IIT (BHU), Varanasi, Proceedings of 5th IRF International Conference, Chennai, 23rd March. 2014, ISBN: 978-93-82702-67-2

[3] Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Rajendran S, Saravanan K, Sobha L, Subbarao K V: Designing a Common POS-Tagset Framework for Indian Languages, The 6th Workshop on Asian language Resources, 2008

[4] Narayan Chaudhary et.al. (2010): ILCI Parts of Speech guidelines document.

[5] Banikanta Kakati, Assamese: Its formation and Development, Lawyers Book stall, 2001

[6] Bipul Roy, Bipul Syam Purkayastha: A Study on Different Part of Speech (POS) Tagging Approaches in Assamese Language, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016

[7] IIIT-Tagset. A Parts-of-Speech tagset for Indian languages http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

[8] Purva S. Dholakia1, M. Mohamed Yoonus: Rule Based Approach for the Transition of Tagsets to Build the POS Annotated Corpus, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014

BIOGRAPHIES

Mr. Bipul Roy is working as Scientist-B in NIELIT Guwahati Kokrajhar Extension Centre. His research interests include Data Mining, Natural Language Processing (NLP), E-learning, Information and communications technology and Computer Applications etc.

Prof. Bipul Syam Purkayastha, is working as Head of the Department, Computer Science in Assam University, Silchar, Assam, India. He is handling various projects in the area of Natural Language Processing, Information Extraction and Retrieval. He has published many national & international papers.