Parts of Speech Tagging in Bengali for MWEs Detection

Md Jaynal Abedin
Department of Computer Science
Assam University, Silchar, Assam, India

Bipul Syam Purkayastha
Department of Computer Science
Assam University, Silchar, Assam, India

ABSTRACT

Part of speech (POS) tagging is the process of assigning the part of speech tag to each and every word in a sentence. In many Natural Language Processing applications such as word sense disambiguation, information retrieval, information processing, parsing, question answering, MWEs detection and machine translation, POS tagging is considered as the one of the basic important tools. Identifying the ambiguities in language lexical items is based on the proper identification of Part of Sspeech (POS) tagging of that language which can enhance the language processing applications in different ways. This paper describes the POS tagset for Multiword Expressions Detection in Bengali (Bangla) which is also very important for many natural language processing (NLP) applications.

Keywords

MWEs, annotation, tagging, noun, verb, adjective, adverb, postposition, part of-speech

1. INTRODUCTION

Due to Bengali (Bangla) language has rich morphological nature, Bangla is a language with a high inflectional system. Inflections include postpositions, number, gender and case markers on nouns, and inflections on verbs include person, tense, aspect, honorific, non-honorific, pejorative, finiteness and non-finiteness. Since syntactical bracketing is a task of shallow processing and size of the tagset is one of the important factors, only postpositions, accusative and possessive case markers on nouns have been incorporated in this tagset. To reflect only these characteristics of morphology, a separate category 'Suffixes' has been included to denote the inflections. When a noun or a pronoun is inflected by a suffix, the base form and inflections are separated by a plus sign (+)[1]. Verbs are categorized according to their form such as finite, non-finite etc.

Multiword Expressions(MWEs) plays an important role in Natural Language Processing because the NLP is concerned with text that may interact with each other. Multiword Expressions (MWEs) have been identified with an increasing amount of interest in the field of computational linguistics and Natural Language Processing (NLP) [2]. Formal definition of Multiword Expression define by **[**3**]** as: Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes, and (b) display lexical, syntactic, semantic, pragmatic or statistical idiomaticity. MWEs are characterized by non-compositionality, non substitutability and non-modifiability **[**4**]**.

We are developing an Annoted corpus for Multiword Expressions (MWEs) detection to improves the efficiency of MWEs detection. Thus, POS tagging help in annotation of Bangla text to form a syntactical Treebank. While tagging, pure lexical category of a word has been preferred to be taken into consideration so far [5;6], because it ensures the consistency in tagging and reduces the confusion involved in manual tagging. It is also helpful for a machine to establish a word-tag relation which leads to efficient machine learning.

2. LITERATURE SURVEY FOR INDIAN LANGUAGES

Compared to Indian languages, foreign languages like English, Arabic and other European languages have many POS taggers [7]. Literature shows that, for Indian languages, POS taggers were developed only in Hindi, Bengali, Panjabi and Dravidian languages.

In comparison to the development in the field NLP, large annotated corpus is slowly growing in Bengali(Bangla), some recent works on experimenting stochastic models [8][9][10] have achieved higher accuracy in automatic POS tagging. It has been shown that the accuracy of the POS tagger can be significantly improved by integrating morphological analyzer, prefix/suffix information, name entity recognizer etc.

3. MOTIVATION FOR THE IDENTIFICATION OF MWEs IN BENGALI

Since many difficulties arise in Bengali POS that motivate us to work on MWEs detection in Bengali. Some examples of MWEs which are difficult in POS tagging are words like কানে লাগা (kany laga) which means 'interesting', কান কাটা (kan kata) which means 'shamless',হাত থাকা (hat taka) which means 'right',উঠন্ত মুলো পত্তনে চেনা যায় (utanto mulo potony chena jaey) which means 'morning shows the day', and so on. Good morphological analyzers, POS taggers, stemmer and annotated corpus etc are not yet available in this task. Bengali is highly versatile language providing one of the most challenging sets of linguistics and rich statistical features resulting in Complex and long word formation. In spite of other Natural language Processing (NLP) task like Information retrieval, Text summarization and Machine translation etc, in Bengali it is needed to identify MWEs along with their detection and extraction process from different domain.

4. STEPS TO POS TAGGING

The first step towards POS tagging is morphological analysis of the words. For this a Noun Analysis and verb Analysis of the words have been done. Nouns are divided into three paradigms according to their endings, these three paradigms are further classified into two groups depending on the feature ± animate. The suffixes are then classified based on number, postposition and classifier information. Verbs are classified into 6 paradigms based on morphosyntactic alternation of the root. The suffixes are further analysed for person and honourofic information [11]. Further steps includes identification of words and their orthographic forms, Analysis of words, morphological structures and their formation, syntactic (grammatical) functions in a sentence, determination of grammatical roles, semantic roles in the sentences, and final verification and validation of the tags that will be assigned to the sentence level.

5. BENGALI TAGSET SUMMARY

We are presenting the list of Bangali tagset which are prominently used in Bengali language for Natural language processing applications.

Table 1. Bengali (Bangla) Tagset

Sl no	Tag Description	Level	Examples
1	Common Noun	N_NNC	বালক (bālak), শহর(śahar), কথা (kathā), মানুষ (Man),
2	Proper Noun	N_NNP	করিম(Karim),দিল্লি(Delhi)
3	Material Noun	N_NNM	কলম (kalam) pen
4	*Nloc noun*	*N_NST*	*উপরে* (upare)
5	Temporal Noun	*N_NNT*	গতকাল (yesterday), আজ (today), এখন (now)
6	Verb root	*N_NNV*	গোসল করা (taking bath), পান করা (drink)
7	Locative noun	*N_NNL*	উপর (up), নিচে (down), আগে (front)
8	Question locative noun	*N_QNL*	কোথায় (where), যেখানে (relative ‘where’)
9	Question temporal noun	N_QNT	কখন (when), যখন (relative ‘when’)
10	Collective noun	N_NNL	দল (dal) ‘party’
11	Abstract noun	N_NNA	ভয়(bhay) ‘fear’
12	Verbal noun	N_NNV	গ্রহণ (grahaṇ) taking, নাইস্ (nice) ভিতেরে(bhitare)
13	Pronoun	PR	আমি (āmi), তুমি (tumi),সে(se), তারা (tār ā), তু ই (tui), etc.
14	Personal Pronoun	PRP	আমি (āmi),সে (se) তু (tumi), আমরা (āmrā**)**
15	Reflexive pronoun	PRF	নিজেকে(nijeke)
16	Relative pronoun	PRL	যে(ýe), যারা (ýārā), যাদের (ýā der), যাকে (ýāke)
17	Reciprocal pronoun	PRC	পরস্পর(paraspar)
18	Wh-word Pronoun	PRQ	কে (ke), কাকে (kāke), কারা (kā rā), কাদের (kāder)
19	Question Pronoun	QPR	কে (who), কারা (plural ‘who’), যে (relative ‘who’)
20	Demonstrative	DM	যে(ýe),এই (ei),ওই (oi), তাই (tāi), etc.
21	Deictic Demonstrative	DMD	এ(e),এই (ei),সে (se),সিই (sei),ও (o),ওই (o
22	Relative demonstrative	DMR	যে(ýe),যেই (ýei) যাহা (ýāhā), যা (ýā)
23	Wh-word demonstrative	DMQ	কানো (kano), কোনা(kona)
24	Finite Verb	FV	করিছ (karchi), করতাম (kartā m),গেলা (gela),যাবে (ýābe),etc
25	Non-Finite Verb	NFV	করলে (karle), করতে (karte),গে লে (gele), গিয়া(giye), etc.
26	Non finite perfective verb	VBT	করা (doing), করানো(causative ‘doing’), পড়া (reading)
27	Subjunctive verb	VBC	করেল (if done) ,
28	Auxiliary Verb	VBX	করে ফেললাম/VBX (have done), হেসে উঠলো/VBX (burst into laughter)
29	Finite Existential	VBE	হয় (be), হবে (will be)
30	Nonfinite Existential	VBEF	হেত (to be)
31	Adjective simple	AD	ভাল (bhāla), মন্দ(manda), সুন্দর(sundar)(beautiful), সাদা (sādā), লাল (red),শ্রেষ্ট(best), শ্রেষ্টতম (the best)etc
32	Verb root adjective	JJV	লাল/JJV হওয়া/VBM (to redden), দুর্বল /JJV হওয়া/VBM (to weaken)
33	Question Adjective	QJJ	কেমন (how), যেমন (relative ‘how’) ,
34	Adverb	AV	হঠাত্ (haṭhāt), বাবদ (bābad), কারণে (kāraṇe), etc
35	Question Adverb	QRB	কেন (why), কিভাবে(how),
36	Postposition	PP	পের(pare), কাছে (kāche), আগে (āge), দারা (by), থেকে(from), জন্য (for), চাইতে (than)
37	Conjunction	CN	তেব (tabe), যদি (ýadi) নইলে (n aile), যাতে (ýāte), etc.
38	Coordinating Conjunction	CC	এবং (and), অথবা (or), নতুবা (nor)
39	Compound coordinating	CCC	না/CCC হয়/CC(neither) **Conjunction Sub types**
40	Suspecion Conjunction	CN	যদি (if), পাছে (if)
41	Eternal joining Conjunction	CET	যেমন/CET ... তেমন/CET (like … like), যখন/CET ... তখন/CET (when … then)
42	Subordinating Comjunction	CS	যে (Complementizer ‘that’),

			এইজন্য (for this)
43	Compound Coordinating Conjunction	**CSC**	তাই/CSC বেল/CS (that's why), এই/CSC কারেণ/CS (for this reason)
44	Interjection	UH	ওহ! (oh!), হায়! (alas!)
45	Indeclinable	IN	কিন্তু (kintu),অথবা (athabā), ব রং (baraṃ), আর(ār), etc
46	Particle	PT	ই (i),ও(o),তা (to),না (nā), নে (ne), নি (ni), etc.
47	Question Particle	QPT	কি (question particle)
48	Quantifier	QT	এক (ek),দুই (dui),প্রথম (pratha m),পয়লা (paylā), etc.
49	Reduplication	RD	চা টা(cha ta),বেন বেন (bane bane), কত কত (kata kata), যেযে(ýe ýe), etc.
50	Foreign Word	FW	যেকোন বিদেশী শব্দ (any foreign word)
51	Postpositional Suffices	SFON	এ, য়, তে
52	Accusative postposition	SFAC	কে, রে, এরে, দিগেরে
53	Possessive postposition	SF$	এর, দের
54	Punctuation Marks	PN	., : ; / …, !, ? ( ), [ ], { }, etc.1 5 Others [OR] Mathematical symbols, +, , x, >, <, $, #, @, ^, &, * etc.

6. SOME EXAMPLES OF POS WITH CORRESPONDING TAG

1. Part of Speech: Compound Common Noun
Tag: NNC
Category: Noun
Examples:
ভারতের/NNP+SF$প্রতোকটি/DMজেলায়/NN+SFONরয়েছেন/VB একজন/QFNUM জেলা/NNC প্রশাসক/NN
"There is one district commissioner at each of the district of India"
বিষয়টি/NN স্বরাষ্ট**/NNC** মন্ত্রালয়ে**/NN+SFON** পেশ/NNV করা/VBM হয়েছে/VBE
"The matter has been submitted to home ministry"
2. Part of Speech: Proper Noun
Tag: NNP
Category: Noun
Example: করিম /NNP একজন/QFNUM জুদ্ধা/NN
"Karim is a warrior"
3. Part of Speech: Compound Proper Noun
Tag: NNPC
Category: Noun
Example:
কাজী**/NNPC** নজরুল**/NNPC** ইসলাম/NNP
"Kazi Nazrul Islam"
4. Part Of Speech: Nominal Verb Root
Tag: NNV
Category: Noun
Examples: সে/PRP গোসল**/NNV** করেছে/VB
"He has taken a bath"
আমি/PRP এখন/NNT চা/NN পান**/NNV** করিতেছি/VB
"I am now taking tea"
5. Part Of Speech: Question Adjective
Tag: QJJ
Category: Adjective
Examples:
আজ/NNT আবহাওয়া/NN অতো/RB সুন্দর/JJ নয়/VB যেমন**/QJJ** আমি/PRP ভেবেছিলাম/VB
"Today's weather is not as that much beautiful as I thought"

7. CONCLUSION

In this paper, we have presented details tagset for Bengali (Bangla) language which is helpful for different level of sentence analysis like Morphological analysis, sentence parsing and level of word selection etc. Finally, it should be stated that the cited tagset can help to build a large Tag Corpus in Bengali language and examples given here can be made more explicit for sentence tagging, based on this Multiword Extraction and Multiword Detection which can be enhanced for the research work in Natural Language Processing. It is also left open for further discussions and suggestions to promote detailed studies of different syntactic phenomena of Bengali without being bound to some traditional and specific syntactic theories.

8. REFERENCES

[1] Altaf Mahmud, Mumit Khan: Syntactic Part of Speech Tagging Guidelines for Bangla Text .Center for Research on Bangla Language Processing (CRBLP), BRAC University, Dhaka, Bangladesh.

[2] Rayson, P., Piao, S., Sharoff, S., Evert, S. & Moriron, B. V. (2010).Multiword expressions: hard going or plain sailing? Language Resources and Evaluation, vol. 44, pp. 1–5.

[3] K. Papineni, S. Roukos, T. Ward, J. Henderson and F. Reeder. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In Proceedings of Human Language Technology , San Diego, CA,pp.132-137, 2002

[4] Jennifer Brundage, M. Kresse, U. Schwall and A. Storrer.1992. Multiword lexemes:A monolingual and contrastive typology for natural language processing and machine translation. Technical Report 232, Institut fuer Wissensbasierte Systeme, IBM DeutschlandGmbH, Heidelberg.

[5] Bharati, A., Sharma, D. M., Bai, L. and Sangal, R. AnnCorra: Annotation Corpora for POS and Chunk Annotation for *Indian* Languages. Language Technologies Research Centre, IIIT, Hyderabad, December 15, 2006.

[6] A Part of Speech Tagger for Indian Languages (POS Tagger).Workshop on Shallow Parsing in South Asian Languages(SPSAL),Twentieth International Joint Conference on Artificial Intelligence, 2007.

[7] Akshar Bharathi and Prashanth R. Mannem (2007), Introduction to the Shallow Parsing Contest for South

Asian Languages", Language Technologies Research Center, International Institute of Information Technology, Hyderabad, India 500032.

[8] Dandapat, S., Sarkar, S., Basu, A. A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. Transactions on engineering, computing and technology VI ISSN 1305-5313. December 2004.

[9] Dandapat, S., Sarkar, S., Basu, A. Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. Proceedings of the ACL 2007 Demo and Poster Sessions, June 2007, pages 221-224.

[10] Ekbal, A., Haque, R., Bandyopadhyay, S. Maximum Entropy Based Bengali Part of Speech Tagging. Advances in natural language processing and applications research in computing science 33, 2008, pp. 67-78.

[11] D. Chakrabarti: Layered Parts of Speech Tagging for Bangla, Problems of Parsing in Indian Languages.