ISSN No. 0976-5697

Volume 4, No. 9, July-August 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Affix Based Word Classification Method of Assamese Text

Bhairab Sarma
Department of Computer Science
Assam University
Silchar, Assam, India
bhairabs@rediffmail.com

Bipul Shyam Purkayastha
Professor, Department of Computer Science
Assam University
Silchar, Assam, India
bipul_sh@hotmail.com

Abstract: Classification of word is an important activity in Natural Language Processing (NLP) analysis. Word classification as we mean in linguistic is not same as in natural language processing. In NLP, the main objective is Part-of-Speech tagging (POST) which if essential for machine translation and language interpretation. However, in linguistic, words are classified as their applications and representation of meaning in the context of real world. Retrieving contextual meaning in language processing is a very challenging job. Because of sense disambiguation, representation ambiguity and words with multiple meaning, the task POST become very difficult. Assamese is a highly inflected and morphologically rich Indian language. In this study, we attempt to classify words based on its morphological structure. We present a method of classification of Assamese word based on its inflectional features. The classes we have used here may not be complement with POS classification. However it could be method of word clustering during POS with application of other smoothing algorithm like HMM, EM etc. We believe that this method can further be implementing into any other inflectional Indian language processing.

Keywords: Affixes, Contextual meaning, NLP, POST, WSD

I. INTRODUCTION

Natural language processing (NLP) is a study on computational language and Artificial Intelligence (AI). The aim of NLP is to interact with computer through human natural languages. The objective of it is to develop user friendly and user sensitive interfaces to communicate with computer. By user friendly and user sensitive interface, we mean here person with disability in learning, vision and mobility impaired could interact through his natural language[1]. To accomplish this, there are so many sub tasks under NLP paradigm to be analysis; like morphological analysis, machine translation, text summarization, language translation and interpretation, voice recognition and synthesis etc. Among them part-of-speech tagging (POST) is one of the important activity. Here, each token (smallest part of a sentence or a written chunk of text) is assign by a predetermined symbol (called tag) depending its category.

The category of a words is again depends on its contextual meaning. The challenging job in this case is to recover the contextual information of each token based on its position in the context. For structural language, like English, it is not so difficult due to structural conformity. For example: 'Ram gave a flower to Sita'. Structure of the sentence is **[SVO].** If we rearrange the words differently, the sentence cannot represent meaning. However in Assamese or any other Indic language it may possible to represent its proper sense. If we classify this sentence grammatically, we get the following classification: 'Ram/(Proper Noun) gave/(Verb past tense) a/ (determiner or article) to/(Preposition) Sita/ (Proper Noun)'. This classification is purely grammatical in nature.

The objective of POS tagging is to classify each word according to its contextual meaning. Unlike English, Assamese is a highly inflectional free word-order language. Assamese is the native language of Assam, spoken by 30 millions of people from Assam [2] and other Indian north-eastern states. It is an official language of Assam and a schedule language of Indian constitution. Since, Assamese is a free order language; classification of words based on contextual information is very difficult. Moreover Assamese words are highly inflected and compound in nature which result more sensible to cause sense ambiguity.

There are so many affix rules available in Assamese grammar. Some affixes are added with noun clause and some are added with verb clause and some are do not inflected. Depending on the category of affixes, we try to develop a methodology for word class classification based on Assamese grammar[3]. In English, all words can be a part-of-speech but in Assamese, there are some differences between pos and word. All POS couldn't be a word directly unlike English. Adding affixes or combining more than one word we can construct a POS in sentences which can further be fragmented into multiple words. The following table-1 shows few examples of word and POS form. Grammatically affix free POS are termed as words.

Table 1-Words vs POS

Word	**POS form**
অসম	অসমীয়াক, অসমীয়াৰে, অসমীয়ালৈ, অসমীয়াৰ
কৰা	কৰবিলৈ, কৰাত, কৰবিলগীয়া, কৰালৈ, কৰাৰ
মা	মাক, মালৈ, মাৰবাবে, মায়ে, মা-জনী

Here we present a method of classification of Assamese word based on its inflection. In the first part of this paper, we outline some morphological rules of Assamese language and then we analyse different affixes that could inflect a word. Next, in the second part we classify each affixes into different cluster and used this cluster to classify words also. We depict our analytical result in third section. Finally we discuss the short coming of our analytical results in the last section of this paper

II. PREVIOUS WORK

In [4 , 9] , using a similar method, a POS tagger has been developed using HMM to retrieve contextual meaning which claim 87% accuracy. Assamese words are formed through three processes: affixation, derivation and compounding. Words derive from another word again inflected by tense, gender and number. They developed a new tagset according to their classification. The main drawback of this tagset is the more number of tags used and non-uniformity with other standard tagset. Since, Assamese is not a fully ordered language, HMM is not suitable to recover contextual information. Our approach is similar to [5], as proposed Sharma and *etal*, where words are classified based on affix evidence . Here, authors classed the word by pattern matching with concurrence of some pivot suffixes. Similar approach was explain in [2,6] to identify noun and verb category only from Assamese text. Our approach is slightly different from these previous methods. We follow pure grammatical rule to identify affix category first. For each word, we extract its affixes and follow up affix rules as mentioned previously, which rule its imply to the word. For example: consider the word '/kitApbor', we find here the suffix 'bor', and as per our suffix rule, 'bor', 'bilAk' are used to represent a group of things or objects. Hence the category of the suffix falls under object groups which result noun category in our classification.

III. ANALYSIS

In our analysis, we broadly classify the words into two categories: inflected and uninflected. Uninflected words are affix free. These uninflected words could be any category. For example: আৰু, নাইবা, দখে◌োন, বতিং etc. these words are never be inflected in any circumstances. However there are very few numbers in this category. Some words are used in sentence without inflection and represent proper meaning. These are termed as root word. To classify this category, we propose to use a dictionary. For example: খা,যা, কৰ, মন, মানুহ etc. Inflected words are again classified into two groups, noun group and verb group. A noun word is inflected by nominative affixes and a verb word is inflected by verbal affixes. Some examples of verbal roots are: কৰ/do, পঢ়/read, খা/eat, যা/go, দ/give, শ◌োা/sleep, নাচ্/dance, শকি/learn, লখি/write, গা/sing, ৰান্ধ/cook, দ◌ৌৰ/run etc.

Further classification is done based on context and application of words as per NLP requirement. These root verbs can be use in sentences either independently or generating a new word by inflection. For example:

a. 'কতিাপ পঢ়'/ Read the book.
b. 'কতিাপ পঢ়াট◌োা ভাল অভ্যাস'/Reading is a good habit.

In the first sentence, 'পঢ়' is an independent root verb can be a word as well as a POS, however in the second sentence, 'পঢ়াট◌োা' is a new word form by inflection, of type noun. Hence adding some tense, aspect and modality (TAM) one category of root can represents different meaning and can falls in different classes. In table-2 some form of new root with inflection are given below.

Table 2-Root with Affix

Verbal Root+ Affix	New Verbal root
খা +উৰা	খুৰা / to eat
কৰ্ +আ	কৰা/ to do
পঢ়্ +ওঁ	পঢ়◌োা / read
শকি্ +আ	শকিা / to learn

All Kinship terms [7] with inflections are considered as noun group. Although we do not use any dictionary for any purposes here, to increase accuracy level we suggest applying dictionary specifically for kinship terms. Kinship terms may or may not takes forms its inflections directly. Some examples are given in table-3 below.

Table 3- Inflectional kinship terms

Noun Root +Affix	Noun Class
মা + জনী	মা-জনী
মামা + বে◌োৰ	মামাবে◌োৰ
দউেতা + সকল	দউেতাসকল
ভাই + হঁত	ভাইহঁত
শহুৰ+ক	শহুৰক
খুৰা+ই	খুৰাই

IV. AFFIX CLASSIFICATION

In Assamese, affixes are classified into four different categories based on their roles played in sentences. These are:

a. উপসৰ্গ (***Prefixes)*** **–** prefixes takes place in front of the root and represent a different meaning of the word in context. These are: প্ৰ, পৰা, অপ, সম, নি, অব, অনু, নৰি, দুৰ, বি, অধ, সু, উ, পৰি, প্ৰতি, অভি, অতি, অপি, উপ, আ । For example: পৰা + জয়/(win) পৰাজয়/ (defeat), অপ + মান/ (respect) অপমান/ (insult) etc.

b. অনুসৰ্গ (***Suffixes)***- suffixes inflected a word placing after the root to express clarity, excuse, suspension and depth of meaning etc. Some examples are ও, ক◌োা, লকৈে, পৰা, এই, ইলি, ইলা etc. Few applications are given here as below:
ক'লৈ + ক◌োা ক'লকৈে◌োা / (anywhere)
আমি +ও আমিও/ (we also)
তুমি + ত◌োা তুমিে◌োা / (you are)

c. প্ৰত্যয় (***Pratyaya)***- pratyayas are affixes which create a new category of word irrespective of classes by inflecting its root word. For example: অসম (Noun) + ঈয়া (Pratyaya) অসমীয়া (Adjective-belongs to Assam, Noun- Assamese language). ৰান্ধ (Verb) + অনী(Pratyaya) ৰান্ধনী(Noun-person who cooks food).

Pratyayas are two types which make a word either noun or verb by inflection. Following table shows some examples of two categories of Pratyayas.

Table 4- Pratyayas classification

Noun pratyaya	Verb pratyaya

ই, অক, ঈ, ইনি, ঈয়া, ইয়া, আ, খন, লোক, বন্নাক, ডাল, টো, জন, গৰাকী, অনী	আ, উৱা, ওঁতা, ৰাঁত, ইলি, ইলা, ওলো, ইল, ইলো, অন

d. বিভক্তি ***(Vibhakties)-*** vibhakties are affixes use to reflect the contextual meaning of a noun in sentence according to their position. Consider the following two sentences:

i) 'মই ৰামৰ কথা কছোঁ' /I am talking about Ram.

ii) 'মই ৰামক কথা কছোঁ'/ I am talking to Ram.

In the first sentence 'ৰামৰ-ৰাম+অৰ (vibhakti)' is added which means about Ram where as in the second sentence 'ৰামক-ৰাম+অক' (vibhakti) is added and means to Ram. Hence vibhakties are added to reflect the proper meaning as per the context of the sentence. Vibhakties are also two types: nominal and verbals. Nominal vibhakties are gender, person and case related while verbal vibhakties are only person and gender related. In Assamese, there are seven vibhakties and these are given in below.

Table 5 Vibhakti List

Vibhakties	Applications
অ, এ, ই	মানুহ, মানুহে
ক	মানুহক
ৰে, দি, দ্ৱাৰা	মানুহৰে, মানুহৰদ্ৱাৰা
লৈ	মানুহলৈ
ৰ পৰা	মানুহৰপৰা
ৰ	মানুহৰ
ত	মানুহত

We can classify these vibhakties base on its inflectional rule. Table-6 shows vibhakties classification with applications.

Table 6 Vibhakti Classification

Verbal Vibhakties	Nominal Vibhakties
ওঁ - খা+ওঁ=খাওঁ/ (I) eat ইম–কৰ+ইম=কৰিমি/(I will) do আ-পঢ়+আ=পঢ়া/(you) read এ-পঢ়+এ=পঢ়ে/(he/she) reads অক-পঢ়+অক=পঢ়ক/(you please) read	ৰ-দউতা+ৰ=দউতাৰ/ (my) father ৱা-দউতা+ৱা=দউতাৱা/(your) father ক-দউতা+ক=দউতাক/ (his/her) father এৱা-ভনী+এৱা=ভনীয়ৱো/ (your) sister এৱ(য়েৱে')-ভনী+এৱ=ভনীয়ৱে'/ (your) sister এক(য়েকে)-শাহু+য়েকে=শাহুয়েকে/ (his/her) mother in law

V. EXPERIMENTAL RESULT

To experiment our approach, we select some text from the Assamese daily news paper 'Dainik Sambad' Epaper version. We developed a module in computer system called 'SMART TOKENIZER[8]' to tokenize each part of sentences from the text. We categorize our samples into five different domain like political, science news, local news etc. We develop a system to extract affixes and in implement here. The same text is again experimented manually using our approach. First we segregate the inflectional word list from un-inflectional and use to experiment our approach. From the second list, we separate the root word that belongs to noun, verb and other category using a dictionary. Our analytical point of view is the inflected word list. Next, from the inflection list we extract the suffixes by comparing with a suffix list. Depending on the suffix category, we class the word accordingly. Next we do the same job manually on the same text. The analysis is given in below:

Total number of words collected: 11510

Table 7 Word analysis

Category of news	Uninflected	Inflected	Total
Political	590	680	1270
Science	670	800	1470
Economics	860	1020	1880
Local news	3150	2540	5690
Sports	570	630	1200
Total words	5840	5670	11510

The number of verb, noun and other classes classified by our system as describe in stage two is given in the following table:

Table 8 Classification by System

Category of news	Total Inflected word	Noun	Verb	Other Class
Political	680	490	186	4
Science	800	576	220	4
Economics	1020	735	276	9
Local news	630	453	170	7
Sports	2540	1830	709	1

Next we run our approach manually in the same text. During analysis we used a small dictionary also to increase accuracy. The total number of inflected word as depicted in table 9 is slightly varied due to inclusion of some roots ending with some suffix words as inflected. Our new analysis data are shorn in table 9. Manual experiment we get

Table 9 Manual experimental result

Category of news	Total Inflected word	Noun	Verb	Other Class
Political	850	498	290	62
Science	766	512	194	60
Economics	1009	690	270	49
Local news	622	445	165	12
Sports	2531	1801	698	32

VI. CONCLUSION AND RECOMENDATIONS

From the tables 8 and 9, we observe significant increased in the figures of third category words. Because of dictionary application, noun and verb classes are decreases which increases third category due to blind inclusion of root words into inflectional group. The percentage of accuracy level of our approach is given in table 10.

Form the accuracy table we observed that system has identify more number of inflected word in noun and verb category but less number of other category. This is because of in Assamese there are many word ending with affix word but they are actually not includes in affixes. They itself is an indecomposable part of the word. For example, in the word 'কঠোৰ', ending with 'ৰ' but it is not an inflected word. In

our system, there is a probability of consider this word as inflectional however in manual testing we take it as another category word probably falls in Adjective group. Further like kinship term, name entities may consider as either un-inflectional or inflectional word in our system due to some of their inflectional behavior. The accuracy level of the system can further be increase by analyzing name entity and binary words separately.

Table 10 Accuracy analysis table

Category of news	**%Noun**			**% Verb**			**% other**		
	System	Manual	Accuracy	System	Manual	Accuracy	System	Manual	Accuracy
Political	72.06	58.59	13.47	27.35	34.12	6.77	0.58	7.29	6.71
Science	72	66.84	5.16	27.5	25.33	2.17	0.5	7.83	7.33
Economics	72.06	68.38	3.68	27.06	26.76	0.3	0.39	4.86	4.47
Local news	71.9	71.54	0.36	26.98	26.53	0.45	1.43	1.93	0.5
Sports	72.05	71.16	0.89	27.91	27.58	0.33	0.28	1.26	0.98
Average	72.014	67.302	4.712	27.36	28.064	2.004	0.636	4.634	3.998

VII. REFERENCES

[1]. Bhairab Sarma & B. S. Purkayastha, "Could NLP Ensure good governance: some Issues and Challenges?", published on a seminar book on Good Governance organized by DDM College, Khowai, West Tripura on 7th December, 2012.

[2]. Navanath Saharia, Utpal Sharma, & Jugal Kalita, "A Suffix-based Noun and Verb Classifier for an Inflectional Language", 2010 International Conference on Asian Language Processing, IEE Computer Society. pages: 19-22

[3]. Dr. Golok Chandra Goswami, "Asamiya Byakaran Prabesh" , fifth edition, Bina Library, Guwahati, Assam

[4]. Navnath Sahariya, Utpal Sharma, Dhruba Jyouti Das and Jugal Kalita, "part of speech tagger for assamese text", published in the proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 33–36, Suntec, Singapore, 4 August 2009.

[5]. Utpal Sharma, Jugal Kalita & Rajib Das, "Classification of Words Based on Affix Evidence", pdf file downloaded from Web link: www.researchgate.net/publication/... Classification_ of_words_based. Downloaded on: 16th March 2013

[6]. Navanath Saharia, Utpal Sharma , Chandan Kalita , "An Extractive Approach of Text Summarization of Assamese using WordNet", available at: www.tezu.ernet.in/~ nlp/paper/gwc_12_word.pdf, downloaded on 16th March, 2013

[7]. Prajadhip Sinha, Bhairab Sarma & Bipul Shyam Purkayastha, "Kinship Terms in Nepali Language and its Morphology", published in International Journal of Computer Applications (IJCA): http://www.ijcaonline.org, Vol. 58- Issue No. 9 (November, 2012), ISSN: 0975-8887, page:43-49

[8]. Bhairab Sarma & Dr. Bipul Shyam Purkayastha, "A Practical Tokenizer for Part of Speech Tagging of English Text", published in International Journal of Research in Computing & Management(IJRCM): http://www.irjcm.com, Vol. 2 (2012), Issue No. 10 (October), ISSN: 2231-5756, Page no: 69-71

[9]. M Mohamed Yoonus & Sameer Sinha, A Hybrid POS Tagger for Indian Languages; 12 Indian Language using Hybrid Approach in 9 Sept, 2011. Available at: www.languageindia.com