A Light Weight Stemmer in Kokborok

Braja Gopal Patra, Khumbar Debbarma, Swapan Debbarma
Department of Computer Science and Engineering
NIT Agartala, India
brajagopal.cse@gmail.com, khum_10jan@yahoo.co.in, swapanxavier@gmail.com

Dipankar Das, Amitava Das, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University, Kolkata, India
dipankar.dipnil2005@gmail.com, amitava.santu@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

Started from the very beginning, Stemming has been playing significant roles in several Natural Language Processing Applications such as information retrieval (IR), machine translation (MT), morph analysis and deciding the part of speech (POS). Several stemmers have been developed for a large number of languages including Indian languages; however no work has been done in Kokborok, a native language of Tripura. In this paper, we have designed a simple rule based stemmer for Kokborok using an affix stripping algorithm. The reduction of inflected words to the stem or root form is performed in the stemmer by stripping the affixes and applying boundary rules where needed. The stemming algorithm has been tested using a corpus of 32578 words and out of which 13044 were uniquely found to have an overall accuracy of 80.02% for minimum suffix stripping algorithm and 85.13% for maximum suffix stripping algorithm.

Keywords: Stemming, part of speech (POS), Kokborok, suffix, prefix.

[1. Introduction]

Kokborok, an Indian language is spoken mainly in the states of Tripura, Assam, Manipur and Mizoram in India and in the neighbouring countries of Myanmar and Bangladesh by more than 2.5 million speakers[1]. Kokborok belongs to the Tibeto-Burman (TB) language family. Kokborok shares the genetic features of TB languages that include phonemic tone, widespread stem homophony, subject-object-verb (SOV) word order, agglutinative verb morphology, verb derivational suffixes originating from the semantic bleaching of verbs, duplication or elaboration, evidentiality and emotional attitudes signalled through sentence final particles, aspect rather than tense marking, lack of gender marking and tendency to reduce disyllabic forms to monosyllabic ones. Very specifically, Kokborok has extensive list of suffixes with more limited number of prefixes and different word classes that are formed by affixation of the respective markers. Kokborok is represented either in Roman script or in Bengali script however Bengali script is less preferred as it is difficult to project the actual

[1] http://tripura.nic.in

tonal effect appropriately. The affixes play the most important role in the structure of the language. In Kokborok, the words are formed in three processes called affixation, derivation and compounding .The majority of the roots found in the language are bound and the affixes are the determining factor of the class of the words in the language.

Stemming is the process of splitting the stem or root part of the word with its affixes without doing any morphological analysis [6]. Stemming is generally used for Information Retrieval, but is also applied for other Natural Language Processing Applications (NLP) such as Machine Translation (MT), Morph Analysis and Part of Speech (POS) Tagging etc. To the best of our knowledge, at present, there is no such available stemmer in Kokborok language. Thus, the developed stemmer can also be used for the development of a root dictionary Kokborok.

An affix stripping algorithm is developed for reducing agglutinated Kokborok words to its stem or root. Maximum root words are bound roots. Affixes are attached to the root words to form a complete word. This algorithm strips affixes and check with the stored affixes for a match, if found then strip the affixes.

The paper is organized in the following manner. Section 2 gives a brief discussion about related works, Section 3 details about Kokborok word formation , Section 4 gives the list of prefixes, suffixes and an example of highly agglutinative word, Section 5 gives the idea about how words are stemmed, Section 6 which includes the experiments and evaluation while the conclusion is drawn in Section 7.

[2. Related Work]

Stemming is required for Information Retrieval, Part of Speech Tagging (POS) and Multiword Expression (MWE) etc. Porter stemmer is one of the famous stemmer for English [9]. Porter came up with the idea of forming root words through manipulation of suffixes. So many other stemmers are also present in English [2], [8]. Stemmer is used in Information Retrieval systems [5] to improve the performance. Recent study shows that non-native English speakers support the growing use of the Internet[2]. This raises the demand of linguistic resources for languages other than English.

In case of Indian languages, the related works are found in Hindi [10]; in which suffixes are striped off on a longest match basis. Another work in Carlos et al., 2009 [1] can be seen where stemmer is used in extraction of lexicon of stems and root word-forms from raw text corpus. On the other hand, a stemming work has been carried out for Bengali [11]. Among all other languages, Manipuri is quite similar to Kokborok as both of the languages fall under the Sino Tibetan language family. A Manipuri stemmer was developed by K. Nongmeikapam et al., 2011 [7]. In Manipuri, both suffixes and prefixes were stripped out in two separate experiments but without applying any rule. They have achieved 81.50%, precision of 91.36% and f-measure of 86.15% for suffixes and for prefixes 70.10%, precision of 76.99% and f-measure of 73.38%.

Even though works on other languages are reported, so far no work has done on Kokborok language as per authors' knowledge. Kokborok is a highly inflected language, thus needed a new approach for stemming.

[2] http://www.internetworldstats.com/stats.htm

[3. Word structure and construction in Kokborok]

In Kokborok, the words are formed by combining a single root word or multiple root words to which single or multiple affixes are attached. Words in Kokborok are basically constructed by *affixation* and *compounding* as shown in Table. 1. The **root word** is the primary lexical unit of a word, and of a word family (root is then called base word), which carries the most significant aspects of semantic content and cannot be reduced into smaller constituents[3]. Content words in nearly all languages contain, and may consist only of root morphemes. However, the term "root" is also used to describe the word without its inflectional endings, but with its lexical endings in place. For example, '*chatters'* has the inflectional root or lemma '*chatter'*, but the lexical root '*chat'*. Inflectional roots are often called stems, and a root in the stricter sense may be thought of as a mono morphemic stem. The traditional definition allows the roots to be either in the form of free morphemes or bound morphemes. In Kokborok generally roots are of two types, *free* and *bound* root. From a statistics we have seen that, out of 32578 words 20289 much of words are bound, 5026 much words are free and rest few compound and others named entity.

Free Roots

The free roots are pure nouns, pronouns, adjectives, and some numerals for example aming (cat), bwrwi (girl). Sometimes, the suffixes are attached to the free root words to signify the number, case, locative, for example amingni (cat's), bwrwirok (girls), kamio (to village) where suffixes 'ni', 'rok', 'o' are used for case, number, location respectively.

[Table 1. Examples of word formed by single or multiple affixation and compounding]

Prefix	Root word	Suffix	Word as written
Bu	Pha (father)		Bupha
	Khai (to do)	di	khaidi
Ma	Thang (to go)	nai	mathangnai
ma+se+ma	Thang	lai+nai	masemathanglainai

[Table 2. Example of word formed by compounding]

Rootword1	Rootword2	Word formed
Ah(fish)	Suri(sword)	Ahsuri(swordfish)
Ma(mother)	Pha(father)	Mapha(mother and father)

Bound roots

[3] http://en.wikipedia.org/wiki/Root_(linguistics)

The Bound root only appears as part of a lengthy word. Verbs in Kokborok always appear in bound form with affixes to give the tense and other information. These are further subdivided as *nominal* and *verbal.*

Nominal bound roots: Nominal bound roots include kinship for example 'ma' (mother), 'pha' (father) to which prefixes 'a' (my), 'bu' (his/her).

Verbal bound roots: Kokborok verbs always occur in bound form to which multiple affixes are added to give the tense, manner of action, for example the word chahdi (eat), chahkha (ate), chahrere (about to eat) has bound root 'Chah' and suffixes 'di', 'kha', 'rere' respectively.

In Kokborok many compound words are found. Compound words are those words which contain more than one root word. Different types of compound words are shown in Table 2. Some compound words are form root word with the addition of prefixes. And the prefix changes according to the person. For example

Achwi-achu (my grandmother and my grandfather) = Ani(my)+ chwi-chu (Grandmother and Grandfather).

Affixes in Kokborok

Kokborok is highly agglutinative and has words which may have more than one affixes attached to the root word or stem. For example

Mathangliyanata(not been able to go) =ma(pref) + thang(RW) + liya(suf) + na(suf) + ta(suf)

Where 'thang' means to go.

Altogether, 91 affixes are there out of which 72 are suffixes and 19 are prefixes. Prefixes are less frequently used as compared to suffixes.

Frequent prefixes that used in Kokborok are bu, bw, ko, kw, ku, jwk, jwla, iri, ki, ke, ka, ma etc.

On the other hand, the frequent suffixes that are used in Kokborok are de, di, drop, bo, ya, na, nai, ni, lai, le, kha, o, khai, rokni, anw, bai etc.

[4. System Design]

The algorithm is designed to remove both multiple suffixes as well as prefixes from the inflected words. It has been observed that the boundary of root words in Kokborok change after addition of suffixes. These boundary changes are dependent on the boundary character and POS of the word to which affixation is taking place. Thus we have added some rules in the algorithm as boundary changes after addition of suffixes.

i.e. kogo = kok(root word)+o(suffix)

rwchabo = rwchap(root word)+o(suffix)

rwchabdi = rwchap(root word)+di(suffix)

kogwi= kok(root word)+ wi (suffix)

The stemming of such words, without applying rules led to meaningless word.

i.e. kogkha→kog + kha

Here "kog" is meaningless word. To avoid this meaningless output after stemming the boundary rules are applied to the boundary character of stemmed word that satisfies the condition. Since not many words exhibit such changes and limitations or constraints of rule less defined the result of stemmer was not very much improved by the incorporation of rules. In a particular word exhibiting boundary changes, it has been observed that only single rule is applicable at a time, simultaneous application of more than one rule is not approved.

In Kokborok, a new approach for stemming is required and several rules are needed to be implemented. In Kokborok the minimum length of root word is two and maximum length of root word is two and maximum length of suffix is ten. Thus, we maintained two separate dictionaries namely prefix and suffix containing the list of prefixes and suffixes. We took a text file containing 32578 numbers of words.

Algorithm:

Stripping -prefixes ()

1. Repeat the step 2 until all the prefixes are removed
2. Read the prefix, if matched then store it in array and decrease the length of string else read another prefix.
3. If length of string >2 then go for suffix stripping, else exit.

Stripping -suffixes ()

1. Repeat the step 2 until all the suffixes are removed
2. Read the largest suffix, if matched then check for rules, then store it in array and decrease the length of string else read another suffix.
3. Exit.

Example: Token=chahnairokno (len=12)

Checking for 0 to 10 from left for prefix i.e. chahnairok no. If prefix found from prefix dictionary strip prefix.

Checking for 0 to 10 from right i.e. ch ahnairokno. If suffix found from suffix dictionary then strip suffix.

Apply rule (replace the last character of stem word to k or p if it is g or b in case the suffix is 'o' or 'wi').

Output: stem+ suffix

Chah+nai+rok+no.

[5. Experiment and Result Evaluation]

The Indian languages are very resource constrained and less computerized to English. A very limited corpus was available as no work has been earlier carried out in Kokborok. The experiments of the systems have been conducted on the corpus collected from Kokborok story books and the holy Bible. The accuracy has been checked manually after applying the algorithm on the corpus that consists of total 32578 words out of which 13044 words are

unique.

[Table 3. Result of Kokborok stemmer]

	minimum suffix first		maximum suffix first	
	Unique words	Whole words	Unique words	Whole words
Applying rule(accuracy)	82.9%	78.56%	85.5%	82.78%
With rule(error)	17.1%	21.44%	14.5%	17.22%
Without rule(accuracy)	80.4%	82.2%	87.9%	84.32%
Without rule(error)	19.6%	17.8%	12.1%	15.68%

We have calculated the accuracy by applying different approaches such as minimum suffix first and then maximum suffix striping. Table. 3 contains the result for minimum suffix stripping first. i.e. suffix stripping from right side of the word. We also applied these both of the algorithms to the whole corpus as well as for the unique words. In Evaluation of the result, the system for affix stripping (minimum suffix) gives an overall accuracy of 80.02%. In our case the mis-stemming, over-stemming and under-stemming leads to low accuracy of the system. For example,

Mis-stemming: tongo= tonk +o (output)

Desired output: tong+o

Under-stemming: brajno=brajn + o (output)

Desired output: braj+no

Over- stemming: bini(input)=bi+ni(output)

Desired output: bini

Out of the total error, there are 45.2% cases of mis-stemming, 31.42% over-stemming and 23.38% under-stemming. In case of Kokborok we have observed that stripping order affect the result. On stripping the suffix with smallest length first the word is under-stemmed when the minimum suffix is a part of the maximum suffix.

Example: buphangno →buphangn+o (under stemmed)

Here the suffix is ‘no’ but also ‘o’ is a suffix that’s why ‘o’ is stripped first, though it’s not a suffix here leading to under-stemmed output.

Table. 3 contains the result for affix stripping (maximum suffix) gives an overall accuracy of 85.13%. There is no case of under stemming seen as we striped largest suffix first. In this case out of the total error, there are 69.3% mis-stemming and 30.7% over-stemming. For example,

Over- stemming: sumano(input)=suma+no(output)

Desired output: suman+o

Mis-stemming cases are same as above.

[6. Conclusion and Future work]

The experiment results of the designed stemmer was found to be promising, however the stemmer can be made stronger by using larger corpus. This stemmer can be implemented for POS tagger, root word collection from corpus, Machine Translation etc. A better approach can be tried to reduce the case of miss stemming, under stemming and over stemming. More rules can be added to the stemmer, which will improve the accuracy but will substantially increase the computational cost. A mixed approach i.e. combination of minimum suffix and maximum suffix first can be tried later. Further unsupervised learning based on statistical machine translation may be applied to improve the accuracy of the current stemmer.

Most of the North-East Indian languages are similar. It will be interesting applying this stemming algorithm upon those languages or similar technique may be used to develop stemmer for these languages.

[7. Acknowledgements]

We would like to thank Dr Binoy Debbarma of Language Wing, Education Dept., TTAADC, Khumulwng, for the assistance in the creation of rules and finding the affixes.

[References]

[1]. Carlos, C. S., Choudhury, M., Dandapat, S., Large-Coverage Root Lexicon Extraction for Hindi. In Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 121--129. Athens, Greece, 2009.

[2]. Dawson, J., Suffix removal and word conflation. ALL Cbulletin 2(3), pp. 33-46, 1974.

[3]. Debbarma, K., Patra, B.G., Debbarma, S., Kumari, L., Purkayastha, B. S., Morphological Analysis of Kokborok for Universal Networking Language Dictionary.In Proceedings of First International Conference on Recent Advances in Information Technology. Dhanbad, India, 2012.

[4]. Debbarma, B., Debbarma, B., Kokborok Terminology P-I, II, III, English-Kokborok-Bengali.Language Wing, Education Dept., TTAADC, Khumulwng, Tripura

[5]. Frakes, W., Baeva-Tates, R., Information Retrieval, Data Structures and Algorithm (eds). Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1992.

[6]. Islam, Md. Z., Uddin, Md. N., Khan, M., A Light Weight Stemmer for Bengali and Its Use in Spelling Checker. In the Proceedings of First International Conference on Digital Communications and Computer Applications, Irbid, Jordan, 2008.

[7]. Kishorjit, Ng., Salam, B., Romania, M., Chanu, Ng. M., Bandyopadhyay, S.: A Light Weight Manipuri Stemmer. In Proceedings of National Conference on Indian Language, Computing (NCILC). Cochin, India, 2011.

[8]. Krovetz, R., Viewing morphology as an inference process. In 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191-202, 1993.

[9]. Porter, M. F., An Algorithm for Suffix Stripping. Program, 14(3):130-137, 1980.

[10]. Ramanathan, A., Rao, D. D., A Lightweight Stemmer for Hindi. In Proceedings Workshop of Computational Linguistics for South Asian Languages- Expanding Synergies with Europe, EACL: pp. 42-48. Budapest, Hungary, 2003.

[11]. Sarkar, S., Bandyopadhyay, S., Design of a Rule-Based Stemmer for Natural Language Text in Bengali. In Proceeding of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 65-72. Hyderabad, India, 2008.