International Journal of Language and Linguistics
2016; 4(2): 74-79
http://www.sciencepublishinggroup.com/j/ijll
doi: 10.11648/j.ijll.20160402.15
ISSN: 2330-0205 (Print); ISSN: 2330-0221 (Online)

The First Step Towards Suffix Stripping of Mising Words Using YASS

Sadiq Hussain[1], Rizwan Rehman[2], G. C. Hazarika[3], J. J. Kuli[4]

[1]Dibrugarh University, Dibrugarh, Assam, India

[2]Centre for Computer Studies, Dibrugarh University, Dibrugarh, Assam, India

[3]Department of Mathematics, Dibrugarh University, Dibrugarh, Assam, India

[4]Department of Ophthalmology, Assam Medical College, Dibrugarh, Assam, India

Email address:

sadiq@dibru.ac.in (S. Hussain), rizwan@dibru.ac.in (R. Rehman), gchazarika@dibru.ac.in (G. C. Hazarika), jawaharjyoti11@gmail.com (J. J. Kuli)

To cite this article:

Sadiq Hussain, Rizwan Rehman, G. C. Hazarika, J. J. Kuli. The First Step Towards Suffix Stripping of Mising Words Using YASS. *International Journal of Language and Linguistics*. Vol. 4, No. 2, 2016, pp. 74-79. doi: 10.11648/j.ijll.20160402.15

Received: January 17, 2016; **Accepted**: February 24, 2016; **Published**: March 21, 2016

Abstract: The authors used yet another suffix stripper (YASS) to find out the base words or stems for one of the languages of north-east India called Mising Language. There are over 5, 00,000 speakers in Mising Language. The Roman scripts are used for Mising Language. Mising Agom Kébang is the highest body of the Mising people and is dedicated for the development of Mising literature. The particular suffix remover may be used without in depth knowledge about the language. The authors successfully used the YASS with a F-score of around 87% for finding the stem. In the field of information retrieval, the automatic removals of suffixes are very important. As the mising language does not have a known corpus, the authors created the corpus.

Keywords: Text Mining, Information Retrieval, Suffix Removal, Mising Language, YASS

1. Introduction

Stemming is used in text mining and information retrieval systems to find the root word by reducing variant word forms. Stemming is a common requirement for natural language processing. The main objective of stemming is to find the root word from its derivational and inflected forms. For indexing and searching purpose, stemming is very important [A.G Jivani., 2011]. Although stemming and lemmatizing are used interchangeably, but they are different in nature. For example, the word inflations like *done, does and doing* will map to the stem 'do'. The word 'did' would not map to the stem in case of stemming. But a lemmatizer would do that. The Stemming algorithms may be classified into three categories. They are truncating, statistical and mixed methods. Lovnis and Porters are popular truncating methods stemmers. The mixed methods are corpus based and context sensitive. N-gram, HMM and YASS (Yet Another Suffix Stripper) are statistical stemming algorithms. In this paper, the authors used YASS stemmer for the Mising Language. It was proposed by [P. Majumder et. al., 2007]. As the stemmer does not depend on the linguistic expertise, it is the main advantage of the stemmer. It was tested for English, Bengali and French language datasets by Prasenjit Majumder, Mandar Mitra, Swapan k. Parui, Gobinda Kole, Pabitra Mitra and Kalyankumar Datta [2007]. The mising corpus was created by the authors with 30,000 words derived from the various books published by Mising Agom Kébang and the Mising dictionary [T. Taid, 2010].The hierarchical clustering and distance measures are used for the creation of clusters.

2. Mising People and Their Language

2.1. Mising People

The Mising are Indo-Mongoloid Schedule Tribe of Assam. The Mising is synonymous with Miri, which means

mediator, intermediary, interpreter.[E. AGait,.1905]. According to Census of 2001, the population of Mising is estimated at 5,87,310. The Misings were inhabitants of the hilly ranges that lie between the Subansiri and the Siyang districts of Arunachal Pradesh. They migrated down to the plains of Assam from an area upstream of the Dihong river in search of better economic life before the advent of the Ahom rules in Assam. Since then the Misings have been living mostly along banks of Brahmaputra River and its tributaries. The Mising still speak their own dialect, which is akin to that of Adis of Arunachal Pradesh and possess their traditional ways of living. Originally, they were worshiper of Donyi (Sun) and Polo (Moon), but at present some of them are followers of Mahapurushia Vaishnav Dharma propounded by Srimanta Sankardeva during 15th and 16th centuries A.D.

2.2. Mising Language

The Mising is a Tibeto-Burman language spoken by the Misings. [T. Taid, 1987]. The languages of some other communities of Arunachal Pradesh are more closely related to the Mising Language. Some of the social groups of the Mising community are Oyan, Dambug, Delu, Moying, Pagro, Sayang and Somuang. The groups hardly show any syntactic variations, but in terms of phonological, morphological and lexical context they are divergent. The Mising language has 14 vowels and 15 consonants. [T. Taid, 1987]. The vowels may divided into two groups 7 short and 7 long types. /m/,/n/,/ny/,/ng/ are the four nasals. /s/ and /z/ are the two fricatives in the language. Mising morphemes can be classified into two categories: root and non-roots. The roots may be classified as nouns and noun substitutes, adjectives including numerals and classifiers, verbs and adverbs. [B.R. Prasad, 1991]

3. Literature Review

[Dalwadi Bijal et al, 2014] discussed different stemming algorithms for non-Indian and Indian language, methods of stemming, accuracy and errors. The authors analyzed various methods suitable for Indian languages viz. Longest matched, Take-all-split method, Finite state automata, N-gram, Brute force technique and look up method. The tested Indian languages are Hindi, Gujarati, Malayalam, Marathi, Punjabi and Assamese. The datasets used were online newspaper, magazine, dictionaries, EMILEE corpus. The approach used were rule based, hand-crafted suffixes, suffix stripping and morphotatic rules etc.

[Reinaldo Viana Alvares et al, 2005] presented STEMBR, a stemmer for Brazilian Portuguese language. The stemmer was based on the statistical study of the frequency of the last letter for words found in Brazilian web pages. The stemmer was compared with other stemmer meant for Portuguese. The result proved the efficiency of the stemmer compared to others. The authors used LexWeb Corpus which is a lexical generator for Portuguese language. The corpus size is approximately 130,000 words. The STEMBR model used three modules for every word. The modules were specific cases, suffix reduction and prefix reduction. The authors concluded that STEMBR model is more efficient than STEMP reference model.

[Abhijit Paul et al, 2014] developed an affix removal stemmer for natural language text in Nepali. The stemming system was based on lexical lookup approach. It is started by introducing different types of lexicon and rules to identify the word in the lexicon. The proposed algorithm removed the unnecessary characters after tokenization. As a part of preprocessing steps, it removed punctuation; digit and single character words. The stemmer performance was evaluated over different domains of 1,800 words. The technique showed improvement in the performance over rule based system. Technology Development for Indian Languages (TDIL) datasets were used for testing with 90.48% accuracy.

[Padmaja Sharma et al, 2012] introduced suffix stripping based named entity recognizer in Assamese for location names. NER is an important task for natural language processing. Although in Assamese language, it was a challenging task as it suffered scarcity of resources. As Assamese is an inflectional language which makes the job more difficult. The work reported a suffix stripping approach to identify those roots of words which are location named entities.

[Navanath Saharia et al, 2012] evaluated stemming algorithms with reference to Assamese language. Assamese is Indo-Aryan, morphologically rich and relatively free word form language. They adopted suffix stripping approach with a rule engine which generated all the suffix sequences. They found 82% accuracy with the suffix stripping approach after adding a root word list.

4. Methodology

As per the details proposed at [P. Majumder et. al., 2007], distance functions are used for mapping a pair of strings s and t to a real number r. If the value of r is small, then it is indicated greater similarity between s and t. They defined a set of string distances $\{D_1, D_2, D_3, D_4\}$ for clustering the lexicon. The main purpose of defining the string distances are to penalize an early mismatch and to reward long matching prefixes. The YASS distance measures D_1, D_2, D_3, D_4 are based on a Boolean function P_i. It is defined as below:

$$P_i = \begin{cases} 0 \text{ if } x_i = y_i \ \ 0 \le i \le \min(n, n') \\ 1 \text{ otherwise} \end{cases} \quad (1)$$

If there is a mismatch in the *I* th position of X and Y, the value of P_i is equal to 1. If $X= x_0\ x_1\ \ldots.x_n$ and $Y=y_0\ y_1\ \ldots.\ y_n$' are two strings and are of unequal lengths and the shorter string would be padded with null characters to make the two strings equal, then D_1would be as follows:

$$D_1\ (X,Y) = \sum_{i=0}^{n} \frac{1}{2^i} p_i \quad (2)$$

The D2, D3 and D4 would be as follows:

$$D_2(X,Y)=\frac{1}{m}\times\sum_{i=m}^{n}\frac{1}{2^{i-m}}\ if\ m>0,\infty\ otherwise$$

$$D_3(X,Y)=\frac{n-m+1}{m}\times\sum_{i=m}^{n}\frac{1}{2^{i-m}}\ if\ m>0,\infty\ otherwise$$

$$D_4(X,Y)=\frac{n-m+1}{n+1}\times\sum_{i=m}^{n}\frac{1}{2^{i-m}} \quad (3)$$

In the equations, *m* denotes the position of the first mismatch between X and Y (i.e. $x_0 = y_0$, $x_1=y_1$, ……., $x_{m-1}=y_{m-1}$ but $x_m \# y_m$). The authors considered two mising words gílenbo (go/come out, taking someone with) and gílenbogor (go/come out, taking someone with, hurriedly) to find out D_1, D_2, D_3, and D_4 as follows:

0	1	2	3	4	5	6	7	8	9
g	i	l	e	n	b	o	x	x	x
g	i	l	e	n	b	o	g	o	r

$$D_1=\frac{1}{2^7}+\frac{1}{2^8}+\frac{1}{2^9}=0.0136$$

$$D_2=\frac{1}{7}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{9-7}})=0.25$$

$$D_3=\frac{3}{7}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{9-7}})=0.75$$

$$D_4=\frac{3}{10}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{9-7}})=0.525$$

The authors again considered two mising words gílenbo (go/come out, taking someone with) and gíndíg (a peak in winter) to find out D_1, D_2, D_3, and D_4 as follows:

0	1	2	3	4	5	6
g	í	l	e	n	b	o
g	í	n	d	í	g	x

$$D_1=\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^6}=0.484375$$

$$D_2=\frac{1}{2}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{6-2}})=0.96875$$

$$D_3=\frac{5}{2}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{6-2}})=4.84375$$

$$D_4=\frac{5}{7}\text{ x }(\frac{1}{2^0}+\ldots\ldots+\frac{1}{2^{6-2}})=1.383928$$

The authors considered two pairs of strings (gílenbo, gílenbogor) and (gílenbo, gíndíg). According to D_1, D_2, D_3, D_4, (gílenbo, gíndíg) are farther apart than (gílenbo, gílenbogor). The above example shows that the distance measures are suitable for the purpose of suffix stripping. After that the lexicon clustering comes into the picture. The distance functions are used for clustering words into homogeneous groups. Each group represents an equivalence class having morphological variants of a single stem. In that cluster, the words are stemmed to the central word. So, the centroids are the stems. The method is broadly discussed at [P. Majumder et. al., 2007].

5. Issues Relating to Mising Word Suffix Stripping

Mising is an agglutinative language. There are more than 400 affixes in Mising. [T. Taid, 2010].

(1)The Suffixes in Mising Language: The majority of suffixes in mising language are derivational. The derivational suffixes are used in the following cases:

(i) Derivation of noun from verb roots:

for example

–ré (nominal suffix, denoting remuneration, charge) jo: ‘ to carry’ > jo: ré (charge for carrying something)

Du: - to sit, to live etc. > du: ré (charges for living somewhere e.g. house rent)

(ii) Derivation of adjectives by adding derivational suffix –né to roots that are adjectival in content. E.g.

Botta-/ botté-/ bétté (to be large in size) – né > bottané / botténé / bétténé (large in size)

Ajji (to be small in size) – né> ajji: né (small in size)

(iii) Derivation of adverbs by using mostly suffix –pé e.g.

Ai (to be good) – pé> aipé (good-ly i.e. well)

Ai (to be good)-mang (suffix marking negative)-pé> aima: pé ‘badly’

Derivation of verbs from verbs to modify the meaning of adverb root or a stem – the largest number of derivational suffixes belong to this category.

Lu- ‘to say’ + -kab > lukab ‘to make someone cry by saying something’

- gor > lugor ‘to say something quickly’
- so > luso ‘to speak less’
- jo: > lujo: ‘to be expert in speaking’

In addition to the above, some suffixes not in large number of pleonastic in nature are used in local dialects, e.g.,-ké: í

Aso: pé, du: toké: í ‘silently sit/be’ > Sit Silently.

(2) Blends in Mising language[T. Taid, 2010]: The formation of blends or portmanteau words are as follows:

(i) In the process of blending, many words meaning male and female of animals or birds are used.

Éki: ‘dog’ + abo ‘male (of animal)’ > ki: bo ‘male dog’

Péjab ‘duck’ +abo ‘male (of animal)’ > jabbo ‘drake’

Péjab ‘duck’+ ané ‘female (of animal)’ > jabné ‘duck’

(ii) Names of parts, words connected with such parts etc. of human bodies and names of many things are formed through similar blending. E.g.

Alé (leg) +amíd 'hair on the body' > lémíd 'hair on the leg'

Amig 'eye' +along 'bone' > miglong 'bone just above the eye socket'

(iii) Some blends do not conform to the pattern of second syllables of words forming a blend.

Le: né (raw)+ asi / así (water) > sile: 'plain water'

sanné (dry) +ongo/éngo (fish) > ngosan 'dried fish'

(3). Plural in mising language [T. Taid, 2010]: kídar/ kíddíng are one of the suffixes used for making it plural. E.g.

Menjég (buffalo) + kídar / kíddíng (suffix marking plural) > Menjégkídar/kíddíng 'buffaloes'

(4). The suffix marking the negative is –mang, which is often reduced to –main word-final positions.

Doma 'eat-not'

(5). Word-final short vowels of monosyllabic words of the structure CV are lengthened when they are followed by a suffix beginning with a consonant [T. Taid, 2010], e.g.

Ngo ('I') + {-mang} (suffix, marking negative) > ngo: mang 'not me'

No('you')+{-rung} (suffix, marking emphasis) > no: rung 'it must be you'

Bí('he/she') + {-yé} (an interrogative suffix) > bí: yé? '(Is it) he/she?'

Sé ('this') + {-lang} (an interrogative suffix, expressing doubt) > sé: lang? '(Is it) this?'

(6). Word-final vowels of disyllabic words remain unchanged in length when followed by a suffix.[T. Taid, 2010]. E.g.

Ami ('person') + {-ko} (the second syllable of ako 'one' used as a suffix) > amiko 'one person'

Adi: ('mountain') + {-to} 'suffix indicating a location to the north of the speaker' > adi: to 'there in the mountains to the north'

Sité ('elephant') + {-dé} 'suffix marking the definite or the specific' > sitédé 'the particular elephant'

The following three tables were used to describe for some of the features of the Mising language. The table-1 described personal definitive inflected on person and number. The word used here is father (Ba: bo). The table-2 described inflectional form of the verb Gerto (means to do) with respect to tense and person. Table-3 described some of the suffixes with categories in Mising Language.

Table 1. *Personal definitive is inflected on person and number.*

Person	Singular	Plural
1st	My father	Our father
	Ngok Ba: bo	Ngoluk Ba: bo
2nd	Your father	Your father
	Nok Ba: buké	Nolukké Ba: buluké
3rd	Her father	Their father
	Bík Ba: bo	Bulukké Ba: bo

Table 2. *Some inflectional form of Gerto (to do) verb with respect to tense and person.*

Gerto (to do)	1st Person	2nd Person	3rd Person
Present	Gerdo/ Gerdag	Gerto	Gertoka
Past	Gertobong/ Gerkabong	Gerton/ Gerkan?	Gerton/ Gerkan?
Future	Geryé	Gerrang	Gerrang
Present Perfect	Gerdung	Gerdu: n	Gerdu: n
Past Perfect	Gerka / Gerkabong	Gerkan/Geramkan/ Geramton?	Gerkabon/ Geramkabon?
Causative		Gerrang	Germoto/ Germolang
Future Conditional	Geryéma: tang	Gerrangka	Gerrangka

Table 3. *Example of suffixes with categories in Mising.*

Plural Suffix	Kídí: dé, kídingé, Nolu, Bojéko, Apping, Abarungko etc.	Tani: kídí: dé
Verbal Suffix	Ka: bong, ka: lang,rang, yí etc.	Gíkabong
Classifiers	Dé, nédé, Ako, Aborko, Siddíko, Sé: bí, édébí etc.	Tani: dé
Case Marker	Lok, yém, yé, lokké, lo: pé etc.	Tani: lo: pé

6. Experiments and Discussion

The performance of the algorithm was evaluated based on different domains of literature published by MAK (MISINGAGOM KÉBANG). These domain includes primarily story books (Do: ying), a collection of Mising priestly rhymes (Mising Ni: tom), a collection of biomonthly news bulletin of MAK (Mimang tíkumsunam). The system was evaluated on 30,000 words based on the corpus. From the corpus, three datasets were derived for testing. The following table provides the statistics of the corpus used.

Table 4. *Statistics of Used Corpus.*

Total Words	Total Nouns	Total verbs
30000	10923	1629

The authors used the evaluation metrics for the dataset is precision, recall and F-Measure. They are defined as follows:-

Recall (R) is the ratio of the number of words stemmed by the system and the total number of words used from the corpus. Precision (P) is ratio of the number of correctly

stemmed words and the total number of words used from the corpus. F-measure is the harmonic mean of precision and recall. Thus mathematically,

$$F\text{-measure} = 2 \times (P \times R) / (P + R)$$

Table 5. *Evaluation of the System with different datasets.*

Dataset	Recall	Precision	F-measure
Dataset-I	89.35%	86.95%	89.13%
Dataset-II	89.56%	85.80%	86.82%
Dataset-III	89.78%	86.26%	87.50%

During stemming process, two text files were generated. One was called correct.txt and another was incorrect.txt. If the system correctly stemmed the word, then the root word would be stored in correct.txt otherwise it would store it in incorrect.txt file. The authors could get total number of root words and incorrect words from these two files. For getting correct root word, one matching program was written. The correct root words were those which werepresent in the corpus as well as in the correct.txt file. The unmatched words were transferred to incorrect.txt. The program would increment the correct root word count if the match was found in both the files.The authors analyzed the incorrect.txt file to find the cases where the system failed to generate the root word. The authors demonstrate the some of the cases below where the system fails.

The phoneme [w] is realized when a vowel viz o, o: and u occurring at the end of word or a morpheme needs to be linked to another vowel sound occurring at the beginning of a suffix. There are about a dozen suffixes in the language that begin with vowels, of which the ones involved in this morphophonemic process are {-ong} (suffix denoting 'only') {-a} (vocative suffix) {-a:} (vocative suffix with greater emphasis) {-é: í} (vocative suffix used when someone calling someone from a distance) {-ar} (emphatic suffix expressing the meaning of 'surely') {- é} (nominative or copular suffix) {- ém} (used mostly as an accusative suffix), {- íng} (with the allomorph, another suffix used for emphasis). E.g.:-

No 'you' + {-ong} /{-o:} > nouwong or nouwo: 'only you'

No 'you' + {-ar} > nouwar '(It's) you, for sure'

So 'here' + {- íng}/ {- í:}> souwíng / souwí: 'right here'

O: 'Mother' + {-é: í}> ouwé: í 'hey, mother!'

Ko: 'boy' + {-a} > kouwa 'hey, boy!'

Ro: 'morning' + {- ém} > rouwém 'in the morning'

Su 'these days' + {-ar} > suuwar 'right at present'

Pao 'name of a Mising Clan' + {-é} > Pauwo '(Someone) is a Pao'

As per Mising phonetic rules, only /p,t,k/ and not /b,d,g/ occurs in the word-final positions. E.g.-

Tabap / tabab 'comb' + {- é} (suffix for the copular 'be') > tababé

Tapat /tapad 'leech' + {- é} (suffix for the copular 'be') > tapadé

Kopak /Kopag 'banana' + {- é} (suffix for the copular 'be') > kopagé. [T. Taid, 2010].

7. Error Analysis

Table 6. *Error Analysis for Mising Language using three different datasets.*

Dataset	Over stemming	Under stemming
Dataset-I	1.30%	9.10%
Dataset-II	4.70%	13.70%
Dataset-III	3.90%	11.30%

There are mainly two types of errors in stemming. They are over stemming and under stemming. Over stemming occurs when two words of different words are stemmed to the same root. This is known as false positive. Under stemming is when two words that should be stemmed to the same root are not. This is called false negative. [Paice, 1990] had proved that light stemming reduced the over stemming errors, but increased the under stemming errors. Heavy stemmers reduced the under stemming errors while increasing the over stemming errors.

The following table described the over stemming and under stemming errors in case of Mising language using YASS for three different datasets.

8. Conclusion

The authors used YASS suffix stripper for the Mising Language. The suffix stripping had not done yet for the Mising Language. The authors found that without much linguistic knowledge about the language, YASS performs well. The number of words taken into consideration is 30000. The authors obtained an F-score of around 87%. As this was the first approach for suffix stripping, the authors did not find other works to compare with. The authors hope to do the parts of speech tagging for the mising language in the future.

Acknowledgement

The authors expressed their gratefulness to Prof. Alak Kr. Buragohain, Vice-Chancellor, Dibrugarh University for his inspiring words. The authors also acknowledged Lalit Kumar Panging, Dibrugarh University for his valuable suggestions. The authors remained grateful to Dr. D. Kardong of Dibrugarh University for having a look at the paper and making necessary corrections.

References

[1] Abhijit Paul, Arindam Dey and Bipul Syam Purkayastha, 2014, An Affix Removal Stemmer for Natural Language Text in Nepali, International Journal of Computer Applications, Vol 91(6)

[2] A. G Jivani., 2011. A comparative Study of Stemming Algoritms, Int. J. Comp. Tech. Appl., Vol 2(6)

[3] B. R Prasad., 1991. Mising Grammar, Central Institute of Indian Languages

[4] Dalwadi Bijal and Suthar Sanket, 2014, Overview of Stemming Algorithms for Indian and Non-Indian Languages, International Journal of Computer Science and Information Technologies, Vol. 5(2)

[5] E. A. Gait,. 1905. A History of Assam. Calcutta: Thacker, Spink & Co

[6] Navanath Saharia, U Sharma, J Kalita, 2012, Analysis and evaluation of stemming algorithms: a case study with Assamese, Proceedings of the International Conference on Advances in Computing, Communications and Informatics

[7] Padmaja Sharma, U. Sharma, J. Kalita, 2012, Suffix stripping based NER in Assamese for location names, 2nd National Computational Intelligence and Signal Processing (CISP)

[8] Paice, 1990, Another Stemmer, ACM SIGIR Forum, Vol 24 (3)

[9] Prasenjit Majumder, Mandar Mitra, Swapan k. Parui, Gobinda Kole, Pabitra Mitra and Kalyankumar Datta, 2007, Yass: yet another suffix stripper, ACM Transactions on Information Systems, Volume 25, Issue 4

[10] Reinaldo Viana Alvares, Ana Cristina Bicharra Garcia, Inhaúma Ferraz, 2005, STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language, Progress in Artificial Intelligence, Volume 3808 of the series Lecture Notes in Computer Science

[11] T. Taid, 1987, Linguistics of the Tibeto-Burman Area, Volume 10.1

[12] T. Taid, 2010, A dictionary of the Mising language: with an introduction to Mising phonology and grammar