Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Suffix Stripping Based Verb Stemming for Hindi

Vishal Gupta

UIET, Panjab University

Chandigarh, India

Abstract— Stemming is the technique of doing mapping of various morphological forms of terms to common term called base or stem. We can also call this technique as conflation on the basis of assumption that words with common base normally have same meaning. Stemming is also largely applied in extraction of information for improving performance of extraction. An algorithm of suffix stripping is not dependent on particular table which has words in inflected forms & relations of root form. But, a particular list carrying lesser number of rules is maintained that can give way for stemming, given a word in a particular form as input, to determine root of that word. The objective of any algorithm for stemming is to get root or stem for terms which are absent from dictionary of that language. After performing stemming if stemmed term is found from dictionary of that language, then we can say that this term is a proper genuine term and if this term is absent in dictionary after stemming then that term can be some name or invalid term Example of stemming is to convert terms fished, fishing & fisher i to common stem term fish .This paper discusses suffix stripping technique for Hindi verb stemming.

Keywords— Hindi stemmer, verb stemmer, verb stemming, suffix stripping

I. INTRODUCTION

Stemming [1] is the technique of doing mapping of various morphological forms of terms to common term called base or stem. We can also call this technique as conflation on the basis of assumption that words with common base normally have same meaning. Stemming is also largely applied in extraction of information for improving performance of extraction. An algorithm of suffix stripping is not dependent on particular table which has words in inflected forms & relations of root form. But, a particular list carrying lesser number of rules is maintained that can give way for stemming, given a word in a particular form as input, to determine root of that word. The objective of any algorithm for stemming is to get root or stem for terms which are absent from dictionary of that language. After performing stemming if stemmed term is found from dictionary of that language, then we can say that this term is a proper genuine term and if this term is absent in dictionary after stemming then that term can be some name or invalid term Example of stemming is to convert terms fished, fishing & fisher to common stem term fish .This paper discusses suffix stripping technique for Hindi verb stemming.

II. RALETD WORK FOR INDIAN LANGUAGES

Very less research has been done for stemming in case of Indian languages. Ramanathan et al. (2003) [2] discussed the Hindi stemmer which is lightweight in nature and it has applied a list of manually made Hindi suffixes & then has used this list for largest stripping match. Islam et al. (2007) [3] discussed another stemmer which is light weight in nature for Bengali language & then used this stemmer for checking of spellings in similar manner as suggested by Ramanathan et al. (2003) [2]. Majumder et al. (2007) [4] suggested YASS (Yet Another Suffix Stripper) which is based on statistical technique and applies clustering oriented technique on the basis of distance among strings & this technique does not demand any knowledge of linguistics. Dasgupta et al. (2006) [5] suggested morphological parsing of Bengali language which is unsupervised in nature and it is process of splitting terms into suffixes, prefixes & stems in absence of previous language oriented rules of morpho-phonological and morphotactics. Pandey et al. (2008) [6] suggested Hindi stemmer which is unsupervised in nature on the basis of technique of split all. Majgaonker et al.(2010) [7] suggested stemmer for Marathi which is unsupervised in nature by including three techniques like: i) suffix stripping ii) rule oriented & iii) stripping based on statistics for creation of rules of suffixes. Suba et al. (2011) [8] suggested two types of Gujarati stemmers : a) inflectional lightweight stemming on the basis of hybrid technique and b) derivational heavyweight stemming on basis of rule oriented technique. Gupta et al. (2011) [9] suggested Punjabi stemmer for names & nouns. In this stemmer an effort was done to get root terms of any word in Punjabi and after this the root term was searched in morph related to nouns in Punjabi & names dictionary.

III. HINDI VERB STEMMING

Verb is meant for explaining state, an action or occurrence & doing forming of main component of predicate belonging to a particular sentence. After analysing the websites of popular online Hindi news papers like: www.bhaskar.com, www.jagran.com, www.amarujala.com and www.punjabkesari.in we have developed a Hindi verb suffix list carrying 30 suffixes of Hindi verbs as given by TABLE I. For performing verb stemming of Hindi, we have consulted Hindi Word-Net [10] available on website of IIT Bombay.

TABLE I
hindi verb suffix list

Sr. No.	Hindi Verb Suffix	Verb Example	Hindi Root Verb
1	ता	खाता "Eats" Masculine and Singular	खा Eat
2	ती	खाती "Eats" Feminine and Singular	खा Eat
3	ते	खाते "Eat" Gender: X and Plural	खा Eat
4	या	खाया "Ate" Masculine and Singular	खा Eat
5	ई	खाई "Ate" Feminine and Singular	खा Eat
6	ए	खाए "Ate" Gender: X and Plural	खा Eat
7	एं	खाएं "Ate" Gender: X and Plural	खा Eat
8	ना	खाना "Eat" Masculine and Singular	खा Eat
9	नी	खानी "Eat" Feminine and Singular	खा Eat
10	ने	खाने "Eat" Masculine and Plural	खा Eat
11	एगा	खाएगा "Will Eat" Masculine and Singular	खा Eat
12	एगी	खाएगी "Will Eat" Feminine and Singular	खा Eat
13	एंगे	खाएंगे "Will Eat" Gender: X and Plural	खा Eat
14	ूंगा	पढूंगा "Will Read" Masculine and Singular	पढ Read
15	ूंगी	पढूंगी "Will Read" Feminine and Singular	पढ Read
16	ोगे	पढोगे "Will Read" Masculine and Singular	पढ Read
17	ोगी	पढोगी "Will Read" Feminine and Singular	पढ Read
18	ये	खाये "Ate" Gender: X and Plural	खा Eat
19	ओगी	खाओगी "Will Eat" Masculine and Singular	खा Eat
20	ओगे	खाओगे "Will Ear" Masculine and Singular	खा Eat
21	यी	खायी "Ate" Feminine and Singular	खा Eat
22	ा	पढा "Read" Masculine and Singular	पढ Read
23	ी	पढी "Read" Feminine and Singular	पढ Read
24	ो	पढो "Read" Gender: X and Plural	पढ Read
25	ूं	पढूं "Will Read" Gender: X and Singular	पढ Read
26	े	भागे "Ran" Gander: X and Plural	भाग Run
27	ो	भागो "Run" Gender: X and Plural	भाग Run
28	ेंगे	भागेंगे "Will Run" Gender: X and Plural	भाग Run
29	ेगा	भागेगा"Will Run" Masculine and Singular	भाग Run
30	ेगी	भागेगी "Will Run" Feminine and Singular	भाग Run

Procedure for Hindi Verb Stemming:

StepA: Input text in Hindi.

StepB: Identify boundary of different Hindi words in input text.

StepC: If any word is present in Hindi Word-Net [10] & it is under verb category of Word-Net then that word is Hindi verb.
Else go to StepD:

StepD: If word is not present in Hindi Word-Net then perform its verb stemming as follows:
If Hindi word ends with any of verb suffixes in the set on basis of longest suffix match as follows:
{ ता, ती, ते, या, ई, ए, एं, ना, नी, ने, एगा, एगी, एंगे, ूंगा, ूंगी, ोगे, ोगी, ये, ओगी, ओगे, यी, ा, ी, ो, ूं, े, ो, ेंगे, ेगा, ेगी } Then remove the corresponding longest suffix from end of that word

StepE: Search the stemmed word in Hindi Word-Net again. If it is found then it is a valid verb and go to StepG. Otherwise go to StepF.

StepF: Current Hindi word is not verb and may fall under other categories of noun, adjective, name or unknown word.

StepG: End of Procedure

Procedure Input: खाई, खाओगे, भागे, पढी

Procedure Output: खा, खा, भाग, पढ

IV. RESULTS AND DISCUSSIONS

This procedure of Hindi verb stemming has been tested on 100 news documents from popular Hindi E-news papers: www.bhaskar.com, www.jagran.com, www.amarujala.com and www.punjabkesari.in and it is found that its efficiency is 83.63%. This efficiency of Hindi Verb stemmer is calculated as follows:

Efficiency of Hindi Verb Stemmer= Number of verbs which are correctly stemmed/ Total number of stemmed terms

Errors of 16.37% are due to absence of certain rules of verb stemming for Hindi or due to non fulfillment of certain verb stemming rules or due to absence of certain root verbs in Hindi Word-Net or due to mistakes in syntax while typing the text. This efficiency of Hindi verb stemmer can be improved by adding more suffixes in suffix list of Hindi verbs and by adding more root verbs in the Hindi-Word-Net.

V. CONCLUSIONS

We can conclude that Hindi verb stemmer has applied suffix stripping oriented rule based technique for doing stemming of Hindi verbs. It includes just thirty suffixes for Hindi verbs. So its efficiency can be improved by adding more suffixes in suffix list of Hindi verbs. Very less number of language oriented resources are available for Indian languages. This stemmer will act as basic resource in language research and will be useful in different applications of natural language processing and text mining.

REFERENCES

[1] www.en.wikipedia.org

[2] A. Ramanathan and D. D. Rao, "A Lightweight Stemmer for Hindi," *Workshop on Computational Linguistics for South-Asian Languages*, EACL, 2003.

[3] M. Z. Islam, M. N. Uddin and M. Khan, "A Light Weight Stemmer for Bengali and its Use in Spelling Checker," *Proceeding of 1st International Conference on Digital Comm. and Computer Applications (DCCA07)*, Irbid, Jordan, 2007.

[4] P. Majumder, M. Mitra, S. K. Parui, G. Kole, P. Mitra, and K. Datta, "YASS: Yet Another Suffix Stripper," *Association for Computing Machinery Transactions on Information Systems*, vol. 25, pp. 18-38, 2007.

[5] S. Dasgupta and V. Ng, "Unsupervised Morphological Parsing of Bengali," *Language Resources and Evaluation*, vol. 40, pp. 311-330, 2006.

[6] A. K. Pandey and T. J. Siddiqui, "An Unsupervised Hindi Stemmer with Heuristic Improvements," *In Proceedings of the Second Workshop on Analytics For Noisy Unstructured Text Data*, pp. 99-105, 2008.

[7] M. M. Majgaonker and T. J Siddiqui, "Discovering Suffixes: A Case Study for Marathi Language," *International Journal on Computer Science and Engineering*, vol. 2, pp. 2716-2720, 2010.

[8] K. Suba, D. Jiandani and P. Bhattacharyya, "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati", *In proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011*, Chiang Mai, Thailand, pp. 1-8, 2011.

[9] V. Gupta and G. S. Lehal, "Punjabi Language Stemmer for Nouns and Proper Names," *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011*, Chiang Mai, Thailand, pp. 35-39, 2011.

[10] http://www.cfilt.iitb.ac.in/wordnet/webhwn

